Review of methods for sublethal aquatic toxicity tests relevant to the Canadian metal-mining industry, and design of field validation programs

Prepared under contract for:

the **Aquatic Effects Technology Evaluation Program**

Scientific Authority:
Danielle Rodrigue, M.Sc.
Canadian Centre for Mineral and Energy Technology
Dept. of Natural Resources Canada
Ottawa
(613) 943.8746

by
John B. Sprague, Ph.D.
Sprague Associates Ltd.
474 Old Scott Road, Salt Spring Island, B.C. V8K 2L7

1995 · November

Introduction

A program called Aquatic Effects Technology Evaluation (AETE) is examining the technology for assessing impacts of liquid effluent from metal-mining. This co-operative program among industry and departments of federal and provincial governments is coordinated by the Canadian Centre for Mineral and Energy Technology (CANMET). AETE is evaluating suitable low-cost techniques for environmental monitoring. Three main areas are (1) lethal and sublethal toxicity testing, (2) biological monitoring in receiving waters, and (3) water and sediment monitoring.

The **sublethal toxicity program** of AETE intends to identify methods that are (a) effective for estimating impacts, (b) of least cost, and (c) linked to a field program. The present report on tests of sublethal toxicity was prepared at the request of AETE, under the sublethal program. Subsequent projects planned by AETE include: **(a) a laboratory program** to compare the performance of sublethal tests; and **(b) a field program** to assess the usefulness of sublethal tests for measuring contributions to observed biological effects in waterbodies.

It is evident that the Management Committee of AETE has particular concern that any laboratory test which might be used for a mining effluent, should have good predictive capability for real-life situations in which the effluent is discharged to fresh water.

Format of this report

The chapters of this report cover three topics requested by AETE. The topics are closely related but discrete, so each chapter is self-contained with independent numbering of sections and its own reference list.

Acknowledgments

Comments, advice, and copies of reports from some members of the toxicity committee of AETE were much appreciated. Documents were kindly provided on short notice by non-committee people Gail C. Oberly of Microbics Inc. in Racine, Wisconsin, and Dr. Robert Spehar of the U.S. EPA in Duluth, Minnesota. I thank Dr. Glenn F. Atkinson of Environment Canada for sharing his knowledge of statistical matters.

Table of contents

Chapter page

1. Rating of sublethal aquatic toxicity tests 2

2. Sublethal tests and their validity in predicting effects on communities in the receiving water 38

3. Planning a field validation program 94

Chapter 1.

Rating of sublethal aquatic toxicity tests.

Table of contents. Chapter 1

Summary and recommendations 4

1 Sublethal tests to be considered 5

2 Explanation of criteria 7
- 2.1 Written document 8
  - *2.1.1 The methods document* 8
- 2.2 Relevance 10
  - *2.2.1 Trophic levels* 10
  - *2.2.2 Phylogenetic category* 11
  - *2.2.3 Choosing species suitable for testing mining wastes* 11
  - *2.2.4 Types of organisms required* 12
  - *2.2.5 Levels of organization* 12
  - *2.2.6 Rating system for relevance of tests* 14
  - *2.2.7 A diverse armoury of tests for screening and other special purposes* 15
- 2.3 Technical characteristics of tests 16
- 2.4 Convenience and economy 17
- 2.5 Calculating the total score 19

3 Numerical rating of the tests 19
- 3.1 General comparison 19
- 3.2 Detailed comments on rating 22
  - *3.2.1 Genotoxicity tests* 22
  - *3.2.2 Biochemical tests* 24
  - *3.2.3 Whole-organism bacterial test* 25
  - *3.2.4 Tests with green plants* 25
  - *3.2.5 Tests with invertebrates* 27
  - *3.2.6 Whole-organism tests with fish* 29

References 30

Appendix. Brief descriptions of the sublethal toxicity tests 33

Summary and recommendations

(1) The following eight tests are recommended for further laboratory evaluation, and possibly for field validation and subsequent consideration. Within each type of organism (same line), the test in bold type (if any) is particularly recommended for evaluation.

Microtox chronic test		
Selenastrum microplate	Multi-species algal microplate	Duckweed
Ceriodaphnia	Nematode	
Fathead minnow larval growth	Trout early life-stages	

(2) For the initial screening tests on any given effluent, each type of organism should be used in testing: fish, invertebrate, plant, and micro-organism, and in addition a test of genotoxicity.

(3) Following that, repetitive routine monitoring could be based on a reduced number of tests. These might be the two simplest and most economical ones which were shown to have adequate sensitivity in the screening of a given effluent. The more time-consuming or difficult tests such as the early life-stages of fish and *Ceriodaphnia* life-cycle, might be dropped from the monitoring program of a given effluent, unless they were needed because of greater sensitivity or some other factor.

(4) The nematode test might prove to be a simple assay to represent multi-cellular animals. As well as further evaluation in laboratory and field programs, the nematode test requires definition of standard methods for testing multiple concentrations and determining the ICp.

(5) Fluorescence should be considered as an endpoint in the *Selenastrum* test and in any other algal test that was used, because of this technique's apparent speed, economy, and freedom from interference by turbidity.

1 Sublethal tests to be considered

The objective of this chapter is to assess a pre-selected list of sublethal aquatic toxicity tests, and give initial recommendations on which tests should be subjected to a laboratory program for evaluating performance.

Thirteen sublethal toxicity tests were pre-selected by AETE for consideration in this chapter. They cover a range of organisms and use several different approaches[1]. As mentioned above, this chapter intends to narrow down the list and provide candidates for a project which will evaluate the performance of toxicity tests in the laboratory.

The representation of organisms on the pre-selected list was diverse, and that is appropriate. Grouping by categories, there was a choice of two or three tests in most categories, but only one whole-organism test for bacteria.

Two tests for genotoxicity, using bacteria
Two biochemical tests
One test on performance of a bacterium
Two tests with algae and one with a multi-cellular plant
Two tests with invertebrate animals
Three tests with the young stages of fish

---

[1] All tests considered in this part of the AETE program are sublethal ones, that is, they are designed to measure a non-lethal endpoint such as reproduction or growth. Although mortality need not occur in the tests to reach an intended endpoint, lethal effects might occur and be used as one of several possible endpoints. For example, death of the delicate larval stages of fish might prove to be the most sensitive effect of any observed during exposure of the early stages of the life cycle.
In some AETE documents, the sublethal tests are called "chronic" (= long-lasting with respect to the life cycle of the organism) which is true for only some of the tests; other sublethal tests being considered by AETE are acute (= rapid).

Table 1 lists the tests and shows the variety of endpoints used. It also shows that the tests have varying degrees of refinement in their methodology documents, but all have received greater or lesser use in Canada or the U.S.A.

A brief description of the main features of each test is provided in the Appendix.

Table 1. List of tests selected by AETE for this evaluation.

The list is organized partly in descending order of trophic level, and partly in descending order of the level of biological integration of the actual test endpoint (see section 2.2.5). A code-word is given for each test, for use in other tables of this chapter. The agency or group which produced the written method is indicated, along with the type of document (SOP signifies Standard Operating Procedure), and a citation to the reference list.

Trout early life-stages, growth & survival	**(TrtEl)**	Published method (Environment Canada 1992d)
Fathead minnow, larval survival & growth	**(FhL)**	Published method (Environment Canada 1992b)
Fathead minnow embryo-larval survival & teratogenicity	**(FhEl)**	Published method (U.S. EPA 1994)
Ceriodaphnia, survival & reproduction	**(Cerio)**	Published method (Environment Canada 1992a)
Nematode survival & growth	**(Nmtd)**	Bioquest International Inc., Published in journal (Samoiloff 1990)
Duckweed growth	**(Dkwd)**	Saskatchewan Research Council, SOP (SRC 1995a) Also published methods (APHA *et al.* 1992, ASTM 1993)
Algal growth & reproduction, multi-species microplate	**(Algae)**	Saskatchewan Research Council, SOP (SRC 1995b) Also Swedish govt. manual (Blanck & Björnsäter 1989)
Algal growth & reproduction, single-species microplate	**(Alga)**	Published method (Environment Canada 1992c)
Microtox chronic test, reproduction & light production	**(MicTx)**	Microbics Corporation (Microbics 1995a, b); also Bulich *et al.* (unpub.)
Metallothionein in liver, rainbow trout *	**(MT)**	Environment Canada, preliminary report (Centre Saint-Laurent 1994)
MFO in liver, rainbow trout *	**(MFO)**	Environment Canada, preliminary report (Centre Saint-Laurent 1994)
SOS Chromotest, genotoxicity	**(SOS)**	Draft instructions (EBPI 1995)
Mutatox, genotoxicity	**(MuTx)**	Manual (Microbics 1993)

* Assessed by measuring specific messenger ribonucleic acid (RNA).

2 Explanation of criteria

Four categories of criteria were set up.

Written document
Relevance
Technical items
Convenience and economy

Each category received separate scores, which were combined to yield a total or final score for ranking the tests.

Written document includes the existence of a written set of instructions and the completeness in specifying the methods and conditions. (See Table 2.)

Relevance is of considerable interest to the committee. A primary component is the degree to which toxicity tests on an effluent are meaningful for predicting events in a waterbody receiving that effluent. In part, relevance signifies the least degree of extrapolation from the laboratory findings to the interactions of a real ecosystem. A component is included to rate existing information on field evaluation. Sensitivity of the test is also included, and a component for pertinence to mining waste in particular. (See Table 3.)

Technical items This category attempts to assess some of the items which make for a meritorious, dependable test. It addresses the appropriateness of the test conditions, validity of results, and proper methods of analysis. (See Table 4.)

Convenience and economy of a test includes the amount of effort, equipment, time, difficulty, and cost to carry out the procedure and process the results. Clearly, a rapid, small-scale, low-labour test with simple equipment would be desirable. Availability of apparatus and existing common use in Canada are also included. (See Table 5.)

Scores in the four categories were weighted before combining them. Relevance was assigned considerable importance by giving it 40% of the total score. The other 60% was equally divided among documentation, technical items, and convenience/economy. Thus the ratios in the total score were 1: 2: 1: 1 for *document: relevance: technical items; convenience/cost.*

The weighting of scores was intended to meet the needs and priorities of the Management Committee of AETE. However, if the committee felt that the relative importance of the categories had been misjudged, it could adjust the magnitudes.

2.1 Written document

This category for the "written document" includes the instructions for doing the test, their completeness, and supporting technical information.

2.1.1 The methods document

A written set of instructions must exist before a test can be used in any testing and monitoring program. It is important to have complete descriptions of all phases of the method, from obtaining organisms and to analysis of results.

A range of formality existed in the test documents. Some tests had sketchy manuscripts or typewritten operating procedures from a particular laboratory. The most formal were publications from an agency, overseen by a committee of people familiar with the test. That kind of document is generally superior since committee members can be expected to have noticed and remedied any items that were missing or less-than-optimal. However, a sketchy or informal document is not a signal for rejecting a method -- a more standardized and complete document could be prepared. Environment Canada, for example, has published a number of standard test methods in this decade.

In the ratings, some attention was paid to the formality of the document as shown in the first group of items in the rating scheme for documents (Table 2). A group or agency publication was assigned a score of 3, compared to a score of unity for an informal method.

The completeness and quality of the description counts for much more than the means of printing. Accordingly, 17 of the potential 20 points were assigned to completeness and clarity. The reasons for the choice of items in Table 2 are to some extent self-evident, and they represent the general coverage that is required in a good methods document. The list parallels the items required for reporting in most toxicity test methods of Environment Canada.

The scores assigned to items of Table 2 were for the most part quantal, either the document gained the single point or it did not. The quality or correctness of the item was generally not judged within this category of written document (e.g. there was no regard for whether the method chose the optimum temperature for culturing and holding *Ceriodaphnia*, only for whether a temperature was specified). Only in obvious cases of inappropriate choice was a score of zero assigned (e.g. use of a statistical method now known to be disadvantageous).

If a procedural item was missing, a score of zero was generally assigned. There was an exception to this, however. If a particular procedural item was clearly irrelevant to a test, then the score was allowed for that item rather than leave the method with an undeserved low score. For example, in the Microtox test, the age/size/stage of organism does not have to be specified in the methods or attended to by the investigator. The organism is supplied in a freeze-dried form under appropriate standard condition. A mark would be allowed for this.

Table 2. Criteria for evaluating documents which describe methods of sublethal testing.

For the first group of items ("Type of written document"), a score can be awarded for only one of the three items listed, i.e. the document could get 3, 2, 1, or 0.

Type of written document		
Published, governmental or group consensus		3
Formal document, standard operating procedure of a laboratory		2
Written but informal stand. operating proc., laboratory or company		1
Test substance		
Handling and documentation specified		1
Organisms		
Potential source(s) described		1
Holding/culturing		1
Age/size/stage specified		1
Criteria for judging/accepting the quality of organisms		1
Physico-chemical conditions		
Specifies temperature, light, oxygen, pH, hardness	1	
Required measurements specified		1
Method of exposure		
Number of individuals specified		1
Replication, if any, described	1	
Observations specified		1
Results		
Endpoint(s) specified	1	
Statistical methods specified	1	
Stated criteria for acceptance of results		1
Requirements for use of reference toxicant		1
Editorial, writing		
Clarity of description, ease of understanding		2
Completeness of coverage, background explanations		1
	Potential score	20

2.2 Relevance

If there is a long path of extrapolation from the laboratory test to the real receiving-water, that can be expected to reduce the accuracy of predicted effects, and hence the relevance of the test results. A short path of extrapolation would be best, and that means that the laboratory test should measure functions that are similar to those in the real world, and important in it.

No single test can be expected to be most relevant, to the exclusion of all other tests. The various types of organisms differ in their sensitivity, and more importantly, the various types will change their rank in sensitivity according to the type of toxic substance that is active. The only suitable approach to that situation is to test a "battery" (small spectrum) of organisms, selected to represent the different types in a natural system. This is further discussed below.

2.2.1 Trophic levels

All the organisms in an ecosystem can be classified into a few trophic levels. Trophic signifies "feeding", and type of nutritive strategy is the general means of setting up the categories.

> *Carnivores* or predators. These animals dine on other animals. They could be small or large, for example there are many carnivorous aquatic insects. There is no need here, to divide this category into primary, secondary, or top carnivores. Carnivores are particularly susceptible to bioaccumulative toxicants.
>
> *Omnivores.* Animals which are not choosy in their diet, feeding on plant or animal matter.
>
> *Herbivores* or primary consumers. Animals which graze, for the most part, on living green plants, large or small.
>
> *Primary producers.* Green plants which manufacture organic material using energy from the sun. Includes everything from single-celled algae, to giant kelp in the sea.
>
> *Decomposers.* Microscopic organisms (bacteria, fungi, protozoans) are usually assigned here, although strictly speaking, some might belong to other categories. Some larger animals might serve a similar decomposing function (e.g. caddisflies that shred dead leaves) but they are usually assigned to the omnivore category.

All these categories can overlap somewhat since most creatures are somewhat opportunistic in their diet and life-style. The classification works by using the primary features of a species, and certainly serves useful purposes in thinking about the workings of an aquatic community.

There are other ways of classifying living creatures such as their particular tactics for making a living: grazers, shredders, burrowers, etc. However, the trophic classification outlined above will be satisfactory for the purposes here.

2.2.2 Phylogenetic category

Test species can be rated for their suitability within categories (e.g. "fish A" is easier to maintain than "fish B"). Species cannot be compared and rated between categories with any meaning (e.g. there is little meaning in "Fish C" gives a better endpoint than "alga D").

The list of organisms considered here (Table 1) represents different taxonomic categories. The standard biological classification is somewhat parallel to a trophic grouping: bacteria are often decomposers; plants are the primary producers; many invertebrates are herbivores or omnivores; and the larger vertebrates (fish) are usually omnivores or carnivores. This overlap is convenient since test species may be selected in terms of taxonomy, with some assurance that they will also reflect particular trophic levels. A taxonomic approach will be used hereafter in this review.

Special vulnerability of particular organisms to a given toxicant (e.g. a metal) is more likely to be discovered by using a battery of tests for initial screening. Good modern practice tests a fish, an invertebrate, a plant, and a micro-organism. In the public's view, fish would be "important" to protect, a perception that might be related to sport fishing, use as human food, or presumed sensitivity of trout. In fact, other groups might be more vulnerable. Micro-organisms can be sensitive and are of the utmost importance. An aquatic ecosystem would function without fish predators, but it would grind to a halt without decomposer micro-organisms. All nutrients would become locked up in detritus which never decayed.

2.2.3 Choosing species suitable for testing mining wastes

Crustaceans and some other invertebrates are highly sensitive to metals, and *Ceriodaphnia* is notably sensitive to some mining wastes (section 3.2.5). Algae and most other plants are also sensitive to metals (see section 3.2.4). However, a particular mine-waste might have variant pH as a more toxic quality than metals, and fish or bacteria might be most sensitive. High dissolved minerals would likely affect *Ceriodaphnia* in particular, as would high suspended solids affect this and other filter-feeders. Thus the varied kinds of mining waste would predictably affect different types of organisms, and we are returned to the principle of testing a spectrum of organisms, at least for initial screening of a waste. *There is no species that is especially suitable for testing all mining wastes, any more than there is a "best" species for testing organic chemicals.*

If wastewater from metal mining affected any single type of organism (e.g. algae which are the primary producers for the community, or micro-organisms which decompose), indirect effects could resonate through the entire community. Accordingly, a variety of types of organisms has been judiciously pre-selected by AETE for this review (Table 1).

Tests with different types of organisms *should not* correlate well with each other. The opposite situation is the expected one, justifying a goal of including a spectrum of organisms, in order to adequately assess toxicity of the different types of pollutants which might be found in mining waste.

Tests considered here are suitable for most kinds of wastewaters including mining wastes. In tests with plants, however, there has been a special attempt to make sure that the duckweed and multi-species algal tests are suitable for testing mining waste (SRC 1995c, see section 3.2.4).

2.2.4 Types of organisms required

The first task of this chapter is, then, to rate the tests and test organisms of Table 1, and recommend some of the most desirable or useful ones. In doing so, there must be a balance among the groups, and at least one must be from each of the following four basic categories.

fish invertebrate plant micro-organism

2.2.5 Levels of organization

Another important component of selecting a test method is the level of organization represented by the effect measured. When biologists speak of levels of organization (= levels of integration), they mean levels similar to the following list (which gives some examples).

ecosystem	(broad, could be 100s of km^2, but ± uniform and elements interacting)
community	(a uniform local assemblage of many species, e.g. a marsh)
species	(all individuals of one kind, capable of interbreeding)
population	(group of individuals of the same species in one locality)
individual	(one fish or one water-lily)
organ system	(the digestive system of a mayfly)
tissue level	(muscle tissue)
cellular	(cells of, say, a gonad, or in and among the cells)
intracellular	(the nucleus or a vacuole within a cell or cells)
biochemical	(an enzyme or enzyme system)

There are various versions of the list. It could be divided more finely, or extended upwards to the *biosphere*, or downwards to the components of molecules. The level of effect measured can be almost independent of the species used (e.g. reproduction of an alga or a fish).

The important principle is that one must move downwards in these levels to discover the *cause* or *mechanism* of an observed phenomenon, while one must move upwards to find the *significance* of the phenomenon. For example, the cause of a growth effect among individual fish might be investigated by studying the digestive system, its food intake and its functioning. If one wished to know the significance of the change in growth, one could measure the effect (if any) on the sizes of fish within the population.

A corollary represents a major concern of the toxicity committee of AETE: *it is difficult to measure a change at one level, then use that information to predict the effect at the next higher level.* Other systems of homeostasis are likely at higher levels of integration. It is difficult to predict one level upwards, and it has been labelled "foolish" to attempt a prediction two or three levels upwards[2]. One should not attempt to predict how changed growth rate of fish will affect the characteristics and balance of a community of aquatic organisms!

---

[2] F.E.J. Fry, one of the most famous Canadian aquatic biologists, at a symposium in his honour, October, 1974, in Algonquin Park, Ontario. (*Natura naturans: a symposium on the Fry paradigm.* J. Fish. Res. Bd Canada (1976) 33:298-345)

These principles have profound implications for rating the relevance of toxicity tests. It means that endpoints of tests should be at or near the top of the list given above, for the greatest relevance in predicting what might happen from discharge of an effluent into a real waterbody.

Clearly there are limits to using endpoints at the top of the preceding list (= high levels of biological integration). Ecosystems and communities are the objects intended to be protected by using laboratory tests, but it is not feasible to run community-level toxicity tests for routine monitoring. Scientific experiments at the community level can require months (e.g. artificial streams) or years (e.g. manipulations of lakes in northwestern Ontario)[3].

For widespread use at reasonable cost, about the best that is currently done for toxicity tests at a high level of integration, is to study survival, growth, and/or reproductive success in small captive populations (e.g. ten or so fish in an aquarium, or an algal culture in a test tube). The tests are not really at a "population" level, since the production of progeny is usually measured under favourable conditions (e.g. asymptotic growth for algae). Such lab tests with small captive populations do not assess population changes under the restraints of a real ecosystem, so they would have to be assigned to the whole-organism or "individual" level of organization.

Does this mean that tests at lower levels of organization are without value as measurements of toxicity? Of course not (see section 2.2.7). The intracellular tests for mutagenic activity (Table 1) could show the potential of a pollutant for causing genetic damage, a category of effect which could be of profound significance. Similarly, elevations in mixed-function oxidases and metallothioneins are known to be associated with toxicant exposures, and they might correlate with damage to communities at a given discharge site. That would have to be documented.

The consequence of these principles of biological organization is that higher ratings for relevance must be given to toxicity tests based on whole-organism performance, particularly reproduction, growth, long-term survival, or indices based on such variables. Such a rating system for relevance is outlined in the following section 2.2.6.

---

[3] "Artificial ecosystems", *mesocosms* and *microcosms* are covered in some detail in chapter 2 but might be briefly outlined here. Cosms are simple communities that might be as large as a pond, or as small as an aquarium or a dish. As a generalization, cosms in containers in a laboratory yield results that are variable and difficult to interpret, making them difficult for toxicity validation.

A series of artificial streams is an excellent way to measure effects on a community and evaluate the predictive ability of laboratory tests. State-of-the-art examples have been the trials by U.S. EPA workers on the effects of toxicants on "natural" invertebrate and fish communities in outdoor experimental channels (e.g. pentachlorophenol, Zischke *et al.* 1983). Each of a dozen or so channels, can be constructed to the same specifications, conditions of inflow carefully controlled, and a natural assemblage of organisms allowed to develop (initial numbers of fish might be controlled). Regulated dosing of the streams at different concentrations, with replicates, allows an exceptional field test. Still, such a test requires all or most of a warm season, the work of at least 4 or 5 specialists, and a moderately expensive facility, so only a limited number of toxicants have been evaluated. Such a system can be used for industrial wastes, but for practical reasons has to be built on site, as has been done for pulp mill discharges (Hall *et al.* 1985).

2.2.6 Rating system for relevance of tests

The rating scheme given here for relevance is intended only for this initial, literature-based evaluation. As stated, the objective is to narrow down the list of candidate tests (Table 1), for use in forthcoming practical assessment in the laboratory. Later there will be trials in the field (section 1.1) for further proving of the methods. Each stage will improve knowledge of relevance of the toxicity tests, and will serve to further rate them for suitability.

Table 3. Rating scheme for relevance of sublethal toxicity tests for monitoring.

For each of the four groups of items, a score can be alloted for only one item. The two top levels of organization are in parentheses because they were not found among the tests to be rated.

Level of organization of test *endpoint*	
community	(4)
population	(3.5)
individual	3
organ system	2.5
cellular, or tissue level	2
intracellular	1.5
biochemical	1
Whole-organism endpoint used	
Life-cycle test	2
Reproductive effect	1.5
Growth or equivalent sublethal effect	1
Documented field validation of test	
Much information including mining waste	3
Much information	2
Some information	1
Sensitivity	
Shown good in comparison tests	2
Information for mining effluents	1
Potential maximum score (excludes parenthetical levels of organization)	10

The rating scheme for relevance (Table 3) has a potential maximum score of 3 for level of organization, since none of the tests in the list qualify for the higher levels above "individual organism". The second group of items in Table 3 gives maximum score to full life-cycle tests rather than abbreviated reproductive tests, or growth tests.

The next group of items dealing with field validation might not be adequately covered in this chapter. The contract called for drafting of this evaluation before completing a full review of field validation (chapter 2). The author's judgement on available information was applied here, and generally conforms with findings of chapter 2. Rating the sensitivity of tests (last group in Table 3) made strong use of comparisons in a report by Keddy*et al.* (1992).

2.2.7 A diverse armoury of tests for screening and other special purposes

This report places much emphasis on whole-organism effects, particularly reproduction and/or growth. It seems that these will have the most relevance in developing a *general-purpose* monitoring program, which is the goal of the present project.

In particular the rating scale does not favour, for the present purposes, intracellular and biochemical tests including those for genotoxicity. However, biochemical tests could certainly have an important role in screening, or even in a monitoring program under particular circumstances. Tests of genotoxicity should be included in all initial screening of a wastewater (see next paragraphs).

Screening tests look for expected or unexpected problems. Before starting to discharge a new effluent, it should be standard practice to run a wide range of screening tests at all levels of biological integration, for example genotoxicity, in case it is an unsuspected problem. If there is any indication that some component of the waste might be genotoxic, it is especially important that a test should be run to assess that danger. If initial screening of a wastewater showed a negative result, in most cases there would not be a need for repetitive routine testing for genotoxicity. If a suspicion of genotoxicity were not removed by the initial screening, such a test might be made part of the routine monitoring procedures.

The Mutatox and SOS Chromotest tests are in this special category of assessing potential genotoxicity. As such they are not intended to yield an estimate of an ICp or NOEC, but rather a quantal answer on whether a genotoxic substance is present or not. To indicate this difference in function, the tests have been separated by a line from the other tests in the rating tables. An attempt has been made to rate them in the same manner as the other tests.

Similarly, measurements of metallothionein are known to be indicative of exposure to certain toxicants including metals. Such a test might be an efficient means of monitoring, if an association were shown between MT and harmful levels of metal. Clearly, the Centre Saint-Laurent is developing the MT and MFO tests as potentially efficient, fast, and economical means of monitoring exposure of fish to pollutants.

2.3 Technical characteristics of tests

This section attempts to provide some rating of whether certain items of the method were well designed, or suitably chosen. Ideally all items of method would be rated but it is difficult to do that objectively for many of the items. Some important things are covered in the scheme shown in Table 4.

Table 4. Rating scheme for technical items in methods of sublethal testing.

Species. At least one species named.		1
Background of information on tolerance of test species		1
Amenability of species to laboratory		
Ease of maintaining culture or stock	1	
Ancillary conditions in test		
Appropriate selection	1	
Endpoint		
Defined, objective, recognizable, measurable		1
Graded response		1
Statistical methods justified and explained		1
Within-test precision		
Coefficient of Variation ≤20%		1
Repeatability over time		
Coefficient of Variation ≤30%		1
Background of data with the defined reference toxicant		1
Potential score		10

It must be admitted that scoring on the basis of Table 4 is subjective. Although each item gets a score of either unity or zero, there is subjectivity in deciding whether or not to assign the score, on the basis of reading the method, and results obtained from it. Perhaps it can be said that attempting the rating is, at least, better than not covering the topic. Discussions of some reasons for rating are given later, in the individual rating sections.

Some explanation should be given here. Under the fifth group of items "Endpoint", there is a mark if the test provides a graded effect such as percent increase in weight, or number of young produced. The other alternative would be quantal measurements such as "different / not different" from the control. The graded measurements can be used to calculate an *Inhibiting Concentration* (ICp) such as the IC25. Quantal measurements can only be used to estimate the no-observed-effect, lowest-observed-effect, and threshold-observed-effect concentrations (NOEC, LOEC, TOEC). Point estimates (e.g., Icp) are regarded as a superior type of measurement.

For the item "repeatability over time", the coefficient of variation (CV)≤30% was derived from methods documents of Environment Canada, in which 30% is used as a limit of expected variation for reference toxicants. The value of CV≤20% for within-test precision is arbitrary but believed to be reasonable.

2.4 Convenience and economy

The convenience of doing a test will have no particular relationship to its scientific merit and relevance. At the same time, there is no reason that a scientifically meritorious and appropriate test should not also be convenient to carry out. For example, the remarkable convenience of measurements by fluorescence is discussed in section 3.1 under*Plants*, and in 3.2.4.

In Table 5, under *Speed*, only one of the potential three items (or none of them) would receive a mark, not all items. For example an exposure-period of 4 hours would score 2, but an exposure-period of 7 days would score zero. Similarly under*Person-hours per test*, only one or neither of the items would be scored. Person-hours includes only the average time for carrying out a test and analyzing the results, per sample of effluent. It does not include the time spent in receiving and handling samples and doing any routine chemical tests to describe them.

Under *Cost* in Table 5, "sample size" would receive only one score (or none). Also under*Cost*, the "cold-room" would be such as is often used to maintain temperatures for tanks of cool-water fish; the alternative of appreciable apparatus for direct cooling of water (> 2 HP) would be considered an equivalent cost and would not merit a mark. A bench-top or floor incubator was allowed without scoring penalty. The indicated capital cost is for equipment to run the test, not the laboratory space itself or normal equipment such as analytical balance, pH meter and good compound microscope.

Table 5. Rating scheme for convenience and economy of sublethal toxicity tests.

Speed (exposure-period)		
≤ 4 hours		3
≤ 1 day		2
≤ 4 days		1
Person-hours/test		
≤ 2 hours		2
≤ 6 hours		1
Cost (see text)		
Required sample size ≤ one litre		2
Required sample size ≤ five litres		1
Holding, test do not require cold-room		1
Capital cost ≤ $10,000		1
Operating/per test ≤ $200		1
Simplicity of design and analysis		
Observations only at end		1
Data into computer program, formatted output of analysis	1	
Equipment available for purchase from Canadian sources	1	
Test is now commonly done in Canadian laboratories		
Government		1
Industrial/consultant		1
Used in Environmental Effects Monitoring (Environment Canada) for another industrial group		1
Potential score		16

2.5 Calculating the total score

For any test, there will be four scores with different potential maximum values.

	Raw score	Adjusted score
Written document	20 marks	x 1 = 20
Relevance	10 marks	x 4 = 40
Technical items	10 marks	x 2 = 20
Convenience and economy	16 marks	x 1.25 = 20
		Total = 100

Relevance has been weighted to constitute 40% of the total possible score. The other categories were assigned 20% each. Obviously the AETE committee could adjust the weighting if desired. Similarly, the committee might wish to establish a minimum score in each category, for passing a test towards further consideration. No test method would be expected to yield a perfect score.

3 Numerical rating of the tests

3.1 General comparison

Adjusted scores above 60%, on individual items, have been boxed to emphasize the two groups. The overall ratings (Table 6) tended to split the tests into two groups. On the right hand side of the table, the three tests with fish and the one with *Ceriodaphnia* scored low on convenience and economy, although their total score was moderate to good.

On the left and middle of Table 6 there are tests with low overall scores, resulting mostly from low ratings in relevance and technical items. The low-ranking tests, moving from left to right in the table, were the two on genotoxicity, the two biochemical tests, and the nematode test. Some of these obtained a good or reasonable score in convenience, in contrast to the fish tests.

The two algal tests were above 60% in all categories. The duckweed test almost achieved that status, but scored only 60% in relevance, chiefly because it measures growth not reproduction, and lacks much field validation. The *Ceriodaphnia* method obtained the highest overall score of the thirteen methods, at 88%. Most written documents were assigned reasonable marks. Only the one for SOS Chromotest was below 60%, and that document was considered a first draft by its marketing company.

Generalizations on the ratings are given after Table 6, discussing the tests from left to right as listed in the table. Detailed consideration of ratings is given in the next section (3.2)

Table 6. Overall scores for the thirteen tests evaluated in this chapter. Full names of the rated items are given in Table 2. Raw values are from Tables 7 to 10. Adjusted scores are intended to emphasize relevance of the tests to real aquatic communities which receive metal-mining wastes. Adjusted scores above 60% are boxed for individual items.

3.5

Genotoxicity. The two tests had low scores, especially in relevance, since the scoring system was purposely set up to rank other kinds of tests. The SOS Chromotest also had a very sketchy draft document which gave little or no information beyond the bare test method; that is puzzling since the test has been in existence for a decade. As mentioned elsewhere, one or both of these tests deserve(s) a role in initial screening of effluents, to detect whether genotoxicity exists. Their role is not in long-term monitoring of discharges which are not genotoxic. In view of the role for initial screening, the AETE committee might wish to include one of these tests of genotoxicity for further evaluation in the laboratory; if so the Mutatox test would be a good choice since it achieved a higher rating, and uses the same apparatus as the Microtox test.

Biochemistry. Tests for mixed-function oxidases and metallothionein had methods that were very well described but suffered in relevance because of the position near the bottom of the levels of integration. If these tests were shown, in the future, to have good correlations with changes in affected communities, they might find use in monitoring.

Micro-organism. The Microtox chronic test was the only whole-organism bacterial test evaluated. It is suitable for further appraisal, judging by its reasonable overall score. This is a new test (the recently-printed manual was delivered during the week of these evaluations). It has little comparative testing as yet, and its low score for relevance would presumably improve if rating were done in a year or two. An early reaction from Canadian trials is favourable (personal communication, G. van Aggelen, B.C. Environment, North Vancouver, B.C.).

Plants. The duckweed and the two algal tests lie in a middle range of scores. Their potential usefulness must be considered similar in view of their relative states of development.

A unique fluorescence method with a specific measuring device is used to measure chlorophyll as an endpoint in the microplate multi-species algal test (SRC 1995b). This method should definitely be evaluated for standard use in the *Selenastrum* test as well, since it has many advantages, especially speed and convenience (SRC 1995c). It has adequate sensitivity and avoids a problem of turbidity interfering with readings of optical density, a problem that sometimes requires expensive counting of algal cells. Microscopic counting of an experiment with 96 samples would take 20 hours, coulter counting and compilation would take 6 hours, but fluorescence readings stored to a computer file would take only 2 minutes.

The multi-species algal test suffered somewhat in the scoring because of sparse data on testing. Since this is a new format for an algal test, there is not a great deal of information comparing results with a diversity of other tests, or comparing with findings from field work. Some information is available from Saskatchewan Research Council on toxicity of metals and a selection of mining wastes to the five algal species used in the test, plus duckweed. The multi-species algal test also suffered in the scoring because of a methods document that was rather rough and incomplete. The document from SRC should be considered as the base for a Canadian test method, but interpretation might be assisted by consulting a Swedish report (Blanck & Björnsäter 1989); it covers some gaps of unexplained items in the SRC report, but does have technical errors in analysis of data.

Invertebrates. *Ceriodaphnia* was a clear winner with a score of 88%, compared to only 54% for the Nematode test.

The Nematode test undoubtedly got a lower score than it deserves, largely because of apparent absence of information on relevance and certain items of methodology. This test with minute worms seems remarkably convenient and even more remarkably economical (an adjusted score for convenience of 14 out of 20, third among the thirteen tests). It might be further considered for field validation since lack of information on relevance contributed to its low score.

Fish tests. Among the three tests with fish, the larval growth test with fathead minnows, (Environment Canada 1992b) has a high ranking of 80%, and would seem to be first choice. The U.S. fathead minnow embryo-larval teratogenicity test also got a moderate to good score. The trout early-life-stage test did not achieve a high score, mainly because of low convenience and lack of information about its relevance. However the trout test would have to remain in the list of tests for further consideration, because the fathead minnow should not be used west of the Rocky Mountains, being a non-native fish.

From this rating process, the selection of monitoring tests for further evaluation in the laboratory is listed below. The four categories of organisms are retained as indicated in section 2.2.4. For each type of organism, the favoured choice, if any, is indicated in bold type.

Microtox chronic test

Selenastrum microplate	Multi-species algal microplate	Duckweed
Ceriodaphnia	Nematode	
Fathead minnow larval growth	Trout early life-stages	

3.2 Detailed comments on rating

3.2.1 Genotoxicity tests

Some information can be given here on sensitivity, variation, and cost of the genotoxicity tests, which are recommended for consideration in screening but not in routine monitoring in the metal-mining program. The information below supports the ratings assigned in Tables 9 and 10.

For sensitivity, Willemsen *et al.* (1995) conclude that the SOS Chromotest is less sensitive than the Mutatox. In a comparison which Willemsen *et al.* compiled themselves, 57 substances were tested with Mutatox, 53 of those were also tested by the standard Ames test, and 40 by the SOS Chromotest. They concluded that for "the majority of the compounds the three tests respond similarly." As might be expected, 16 substances had differing results with the three tests. For these, Willemsen *et al.* concluded that "Mutatox appears to have the widest sensitivity spectrum of the three tests." Mutatox tests were positive for 14 of the 16 substances with differing results, the standard Ames test was positive for six, and SOS tests were positive for only one. The general conclusion appears to be that there are similar results for the two tests of genotoxicity being considered in the present chapter.

There is an apparent contradiction of the preceding conclusions in a review of the general literature made by the same authors (Willemsen *et al.* 1995). The SOS test revealed genotoxicity more often than did Mutatox, although the latter usually gave a stronger response. The literature review was largely based on work by Canadians Dutka, Wu, and co-authors. Another study of 14 chemicals showed that Mutatox agreed with the standard Ames test results in 93% of cases, and SOS Chromotest agreed with the Ames test in 88% of cases (Legault *et al.* 1994). Ability to discriminate between known carcinogens and non-carcinogens (i.e. accuracy) was 82% for Mutatox and 64% for SOS Chromotest (the Ames test had only 73% accuracy in these trials). Mutatox and SOS Chromotest were considered of equal sensitivity in detecting low concentrations of chemicals.

For variability, no interlaboratory comparisons of results for Mutatox were found in the literature by Willemsen *et al.* (1995), although tests of the same materials by different authors were in general agreement. The SOS Chromotest has good reproducibility judging by the literature review of Willemsen *et al.* (1995). They cite work showing agreement between authors for most tests of 103 substances, with conflicting findings in only two of the 103.

The operating cost for one Mutatox test has been estimated as about $Can 70, and Willemsen *et al.* (1995) consider it practical. Economy relies on the laboratory already having the special photometer which is also used for the Microtox test. The operating cost for one SOS Chromotest has also been estimated as about $Can 70, and Willemsen*et al.* (1995) regard it as a "cheap, quick and easy genotoxicity assay" which does not require expensive equipment.

Table 7. Detailed ratings of the written methods documents for sublethal tests. The complete wording that describes each rated item is given in Table 2.

7.25

3.2.2 Biochemical tests

The method of the Centre Saint-Laurent evaluates the relative quantity of messenger ribonucleic acids (RNA) which control the generation of metallothionein (MT) and cytochrome P450IA1 (here called MFO for mixed-function oxidase). Measuring the specific RNAs avoids interference by secondary pollutants, in the generation of MT and MFO. Both MT and MFO show increases in fish with exposure to certain toxicants, and MFOs have been extensively studied as indicators of exposure to pulp mill effluents. Metallothioneins respond to metals such as copper and zinc (Dixon and Sprague 1981, Bradley *et al.* 1985) and so could be relevant to metal-mining discharges. There is much scientific research on MTs, but their development in fish is a complicated topic, and the levels attained and maintained are not in simple relationships to the degree of exposure to toxicants, nor to the toxic levels (Hobson and Birge 1989).

Table 8. Detailed ratings for relevance of sublethal tests for predicting effects in freshwater communities.

5+

As concluded above, these biochemical measurements remain to be proven as a routine monitoring tool. They would require field validation. The Centre Saint-Laurent of Environnement Canada considers this development of methods to be experimental at the moment. It is part of a battery of toxicity tests to evaluate the potential risk of pollutants in the aquatic environment. There is certainly a prospective role for MT and MFO as early-warning signals for detecting and assessing exposures to industrial wastes.

3.2.3 Whole-organism bacterial test

The Microtox chronic test is the only one examined in this category. As mentioned in section 3.1, it received a moderate to good rating overall. The chief weakness was lack of information on sensitivity and on field validation, since this is a very new test.

The test is recommended for further evaluation and it should be included in initial monitoring of metal-mining discharges. It would provide the micro-organism component necessary for a balanced assessment at the four recommended trophic/taxonomic levels. The Microbics chronic test seems likely to become extensively used in North America and elsewhere; it shares the automated photometer that serves the very popular acute Microtox test.

Initial evaluation of precision indicates a within-test variation yielding an average CV of 13%, with 90% of CV values ≤21%, according to Microbics (1995a). There do not appear to be data on repeatability, so no rating was given for that quality in Table 9. Sensitivity appears to be similar to that of the *Ceriodaphnia* sublethal test, according to initial testing (Bulich *et al.* unpub.)

3.2.4 Tests with green plants

The scores from this evaluation were not greatly different for the three tests with plants, and all of them are about equally recommended here. All appear to be relevant, small-scale and of reasonable duration. Knowledge and preference of the committee might favour one or the other. In particular the cost, convenience and labour in doing a test might be a deciding factor. With the present methods, the duckweed test is convenient but of slightly lower sensitivity. The *Selenastrum* test can be subject to interference with turbidity, as mentioned above. That can necessitate individual visual counts of algae which is time-consuming and costly; it could elevate the cost from < $200 with an endpoint of optical density to > $700 with visual counts.

Fluorescence should be considered, as recommended above, for measuring the endpoint of all algal tests. That would usually remedy any turbidity problem. Another general recommendation is that there be consideration of wider use of the culture media developed by the Saskatchewan Research Council for tests with duckweed and multiple species of algae. Each medium was designed for use in testing metals, and should produce greater sensitivity compared to media listed by Environment Canada (1992c) or by U.S. test methods.

Mining wastewater has been specifically targeted in design of the duckweed and multi-species algal tests, for example in choosing species and strains of algae and modifying the makeup of the culture medium. A small body of test results with metals and mining effluents has been obtained (SRC 1995c). Duckweed sensitivity was about equal to the least sensitive of five algal species. Within the multi-species test, most comparisons showed that results for the five algal species varied over an order of magnitude, and some spanned two orders of magnitude, in tests with eleven samples of mining wastewater (SRC 1995c). That gain in scope and sensitivity is an advantage in using the different species. Indeed Blanck et al. (1984) stress the importance of testing a battery of algal species because they found an 8- to 30-fold variation in tolerance of metals, and relative sensitivity changed with the toxicant.

For duckweed, sensitivity deserves better documentation (Table 8). Keddy *et al.* (1992) show moderate sensitivity in four comparisons with lethality tests with fish. Field validation is not strong (Table 8). The test method itself appears satisfactory (Table 9) and the convenience and economy are good, ranking as highly as any other test (Table 10). Duckweed has a major advantage for testing minewaters because turbidity does not cause procedural difficulties.

Table 9. Detailed ratings for technical characteristics of methods for sublethal aquatic tests.

5+

The single-species microplate test with ***Selenastrum capricornutum*** has appreciable information available. It proved good to moderate in sensitivity, in comparisons with one or both of two acute bacterial tests and a lethality test with *Daphnia* (summary by Keddy *et al.* 1992). Averaging sensitivities, the algal microplate test ranked 1.4 out of 3 for the 22 comparisons available (calculated by assigning 1 for most sensitive of 3 tests, 2 for second most sensitive, etc.) That summary of comparisons is not very informative, however, since it does not make a comparison with other sublethal tests. A review of sublethal micro-tests by Willemsen *et al.* (1995) gave an overall rating of "sensitive" to this micro-plate test. Their definition of sensitive was "up to an order of magnitude more sensitive than all other tests to at least one compound". They say the test "is very sensitive to metals and oxidizers. It is not very sensitive to organics, except herbicides." The sensitivity to metals would be relevant for testing discharges from metal mining.

Variation of the algal microplate test was summarized from literature sources by Keddy*et al.* (1992) with almost all within-lab CVs being below 20% and most between-lab CVs below 30%. This resulted in the good rating for precision and repeatability in Table 9. Similarly, Willemsen *et al.* (1995) characterized the test as having "good reproducibility" citing literature values for among-laboratory CVs in the regions of 25% and 11 - 41%.

On the deficit side of the algal test, there does not appear to have been any purposeful attempt at field validation, judging by the review of Keddy *et al.* (1992), and hence there is a low score for validation in Table 8. The test is considered rather slow and effort-intensive by Willemsen *et al.* (1995), in their comparison with a selection of shorter-exposure micro-tests.

The multi-species algal microplate test suffers mainly from being in an early stage of development, without a great deal of use to date. The test had a variety of moderate to low scores (Tables 7 to 9). There is no intrinsic reason that it should not achieve a rating similar to the *Selenastrum capricornutum* test which is very similar.

The multi-species algal test, including *Selenastrum* as one of five or more test species, would usually be more sensitive than any test with a single species of alga. AETE should balance the potential for more sensitivity with the additional work in culturing and testing several species. If used as a formal monitoring test, a single endpoint would be desirable and should be defined, rather than allowing the confusion of reporting five or more endpoints.

3.2.5 Tests with invertebrates

Ceriodaphnia rated as an outstanding test, with perfect or near-perfect scores for written method, relevance, and technical items (Table 6). The Environment Canada (1992a) document covered much supplementary information. A low score (10 out of 20) was assigned only for convenience and economy, because the test took more than four days with care and feeding. Some laboratories have had difficulties in culturing and testing (Environment Canada 1992a).

The *Ceriodaphnia* reproductive test is known to be sensitive to metals and many pesticides, and presumably to mining wastewaters. In the comparison by Eco-Research (1991), this test was particularly sensitive to a mining effluent among the eight industrial effluents tested. An effect was measured at 12.5% concentration of the wastewater, whereas the IC50 for an algal test was 38%, and no significant effect was measured for an embryo-larval test with fathead minnow.

Despite some disadvantages, the *Ceriodaphnia* test is one of the most widely used sublethal aquatic toxicity tests in North America, with an extensive data-bank on diverse chemicals and wastes. It is widely used in Canada for pulp and paper discharges, and should be considered for further evaluation in the metal-mining study. Variability of the test is relatively good; four sets of comparisons show that CVs are generally <20% within a laboratory and ≤30% among laboratories (Keddy *et al.* 1992). That accounts for the scores assigned in Table 9.

Table 10. Detailed ratings for items of convenience and economy of sublethal tests.

6

Use of the *Ceriodaphnia* test in field validation is very good. As mentioned above, this test and the fathead minnow larval test have been widely used in the U.S.A. in attempts at validating laboratory predictions of conditions in the field. The ranking of 3 for field validation in Table 8 is therefore warranted, for example there are ten publications cited by Keddy*et al.* (1992, p. 161).

The nematode test with *Panagrellus redivivus* (Samoiloff 1990) deserves further consideration despite its rather low ranking. The test is very small-scale but uses an organized multi-cellular animal. The nematode can be cultured in one-litre jars, apparently with relatively little effort. Tests are run in 2.5-mL autoanalysis cups, with no specific care or inspection during the 4-day exposure. Clearly, the required sample of wastewater would be very small, say 100 mL for routine tests. Further evaluation might indicate that the nematode test was a simple assay to represent multi-cellular animals.

The nematode test might have some difficulties. Some handling and procedural skills would have to be developed to work with the microscopic animals. Some familiarity with culturing would be required as with all organisms. A more detailed procedural manual would be required (see low scoring in Tables 7 and 9). The procedure would have to define standard methods for testing a series of concentrations and estimating the ICp. A compilation should also be made, of tolerance for various standard toxic substances and types of wastewater.

3.2.6 Whole-organism tests with fish

The fathead minnow test of larval survival and growth scored well in three of four categories. The written document is one of two Environment Canada methods which received a perfect score (Table 7). Lowest rating of the larval test was in convenience/economy, largely because it requires a seven-day exposure in a labour-intensive mode for feeding the young fish (Table 10). Accordingly this is not a really cheap test.

The test did well for relevance (Table 8) because a very similar 7-day test has been heavily used in field validation projects by the U.S. EPA and others. Keddy*et al.* (1992) review seven such trials. Sensitivity of the test has been documented as good.

Within-test precision appears to be satisfactory for fathead minnow tests with early life-stages, but repeatability among laboratories seems marginal. For 17 among-laboratory comparisons of 7-day larval tests reviewed by Keddy*et al.* (1992), the mean CV was ≤29.6%, just under the limit of 30% set in Table 9. Among ten laboratories, the CVs varied from 13% for survival of larvae to 52% for weight (API 1988). The test was assigned only half a mark for this among-lab (repeatability) performance.

The fathead minnow embryo-larval survival and teratogenicity test can be expected to deliver similar performance to the larval test. Surprisingly, since larval fish are not weighed, the sensitivity is similar to the larval test. Published papers have argued both ways as to whether growth or mortality is a more sensitive indicator of effect in the early-life-stage tests with fathead minnows. In testing against eight industrial effluents, Eco-Research (1991) described high sensitivity and "ruggedness" (sensitivity to a variety of toxicants) for the fathead minnow embryo-larval test (and the algal growth test with microplates). (The *Ceriodaphnia* reproductive test was "rugged" but somewhat less sensitive because mortality affected measurements of reproduction.)

Field validation would also be similar to the results mentioned above for the larval test. Results of the embryo-larval test were closely correlated with ecological effects in a stream affected by municipal wastewater (Birge *et al.* 1989).

The trout early-life-stage test is similar in scoring to the fathead minnow test, but there is not an extensive data-bank on toxicity results for a standard method, nor is there yet much field validation. Variability of the test does not appear to be well documented. Some laboratories report difficulties in obtaining rainbow trout eggs for the test, in the first half of the year, while other laboratories are apparently able to maintain spawning at all times of year.

References

APHA, AWWA, WEF 1992 [Amer. Public Health Assoc., Amer. Water Works Assoc., and Water Environ. Fed.] Duckweed (proposed). Section 8211, p. 8.39-8-42 *in* Standard methods for the examination of water and wastewater. 18th ed., APHA, Washington, D.C.

API 1988 [American Petroleum Institute]. Fathead minnow 7-day test: round robin study. Intra- and interlaboratory study to determine the reproducibility of the seven-day fathead minnow larval survival and growth test. Amer. Petroleum Inst., Health & Environ. Sci. Dept, Washington, D.C. API Pub. No. 4468.

ASTM 1993. Standard guide for conducting static toxicity tests with *Lemna gibba* G3. Designation E 1415-91. P. 1232-1241 *in* 1993 Annual book of ASTM standards, Vol. 11.04. American Society for Testing and Materials, Philadelphia, Pa.

Birge, W.J., J.A. Black, T.M. Short and A.G. Westerman 1989. A comparative ecological and toxicological investigation of a secondary wastewater treatment plant effluent and its receiving stream. Environ. Toxicol. Chem. 8: 437-450.

Blanck, H., and B. Björnsäter 1989. The algal microtest battery. A manual for routine tests of growth inhibition. Swedish National Chemicals Inspectorate, Science and Technol. Dept., KEMI Rept. no. 3/89, 27 p.

Blanck, H., G. Wallin, and S-Å. Wängberg 1984. Species-dependent variation in algal sensitivity to chemical compounds. Ecotoxicol. Environ. Safety 8: 339-351.

Bradley, R.W., C. DuQuesnay and J.B. Sprague 1985. Acclimation of rainbow trout to zinc: kinetics and mechanism of enhanced tolerance induction. J. Fish. Biol. 27: 367-379.

Bulich, A.A., H. Huynh, and S. Ulitzur unpub. The use of luminescent bacteria for measuring chronic toxicity. Microbics Corp., Carlsbad, Calif. Ms. submitted for publication, 17 p.

Centre Saint-Laurent 1994. Evaluation de la toxicité létale et sublétale avec hépatocytes de truite. Version préliminaire. Environnement Canada, Centre Saint-Laurent, Rept. CPQ400DO, 30 p.

Dixon, D.G. and J.B. Sprague 1981. Copper bioaccumulation and hepatoprotein synthesis during acclimation to copper by juvenile rainbow trout. Aquatic Toxicol. 1: 6981.

EBPI 1995 [Environmental Bio-detection Products Inc.] SOS Chromotest. Standard Operating Procedure. EBPI, Brampton, Ont. First draft, 8 p.

Eco-Research (Canada) Inc. 1991. Evaluation of Canadian and American bioassay procedures used in the assessment of complex wastewaters. Unnumbered report for Canadian, U.S., Québec, and Ontario government agencies, 50 p. + appendices.

Environment Canada 1992a. [Written by D.J. McLeay and J.B. Sprague.] Biological test method: test of reproduction and survival using the cladoceran *Ceriodaphnia dubia.* Environment Canada, Conservation & Protection, Ottawa, Ont. Rept EPS 1/RM/21.

Environment Canada 1992b. [Written by J.B. Sprague and D.J. McLeay.] Biological test method: test of larval growth and survival using fathead minnows. Environment Canada, Conservation & Protection, Ottawa, Ont. Rept EPS 1/RM/22.

Environment Canada 1992c. [Written by D. St-Laurent, G.L. Stephenson, and K.E. Day] Biological test method: microplate growth inhibition test with alga *(Selenastrum capricornutum)*. Environment Canada, Conservation & Protection, Ottawa, Ont. Rept EPS 1/RM/25.

Environment Canada 1992d. [Written by M.R. Gordon, D.J. McLeay and J.B. Sprague.] Biological test method: toxicity tests using early life stages of salmonid fish (rainbow trout, coho salmon, or Atlantic salmon). Environment Canada, Conservation & Protection, Ottawa, Ont. Report EPS 1/RM/28

Hall, T.J., R.K. Haley and D. Liedkie 1985. Effects of biologically treated kraft mill effluent on cold water stream productivity in experimental stream channels -- fourth progress report. NCASI Tech. Bull. 474: 182 p.

Hobson, J.F. and W.J. Birge 1989. Acclimation-induced changes in toxicity and induction of metallothionein-like proteins in the fathead minnow following sublethal exposure to zinc. Environ. Toxicol. Chem. 8: 157-169.

Keddy, C., J.C. Greene and M.A. Bonnell 1992. A review of whole organism bioassays for assessing the quality of soil, freshwater sediment and freshwater in Canada. Prepared for the National Contaminated Sites Remediation Program, CCME Subcommittee on Environmental Quality Criteria for Contaminated Sites. Canadian Council of Ministers of the Environment, Ottawa, Ont. 294 p.

Legault, R., C Blaise, D. Rokosh and R. Chong-Kit 1994. Comparative assessment of the SOS Chromotest kit and the Mutatox test with the *Salmonella* plate incorporation (Ames test) and the fluctuation tests for screening genotoxic agents. Environ. Toxicol. Water Qual. 9: 45-57.

Microbics 1995a. Microtox chronic toxicity test. Microbics Corp., Carlsbad, Calif. 38 p.

Microbics 1995b. Microtox chronic toxicity testing software user's guide. Microbics Corp., Carlsbad, Calif. 69 p.

Microbics 1993. Mutatox manual. Microbics Corp., Carlsbad, Calif. 22 p.

Samoiloff, M. 1990. The nematode toxicity assay using *Panagrellus redivivus.* Toxicity Assessment: An Internat. J. 5: 309-318.

SRC 1995a [Saskatchewan Research Council]. *Lemna minor* toxicity test. SRC Water Quality Lab., Regina. Standard Operating Procedure. SOP 199, 10 p.

SRC 1995b [Saskatchewan Research Council]. Phytoplankton microplate growth test using fluorescence as the endpoint. SRC Water Quality Lab., Regina. Standard Operating Procedure. SOP 204, 23 p.

SRC 1995c [Saskatchewan Research Council]. Development of aquatic plant bioassays for rapid screening and interpretive risk assessments of metal mining wastewaters. SRC, Tech. Univ. Denmark, & Univ. Saskatchewan, Regina. SRC Pub. E-2100-2-C-95, 191 p.

U.S. EPA 1994. Fathead minnow, *Pimephales promelas*, embryo-larval survival and teratogenicity test method *1001.0.* Section 12, p. 114-143 *in* Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms. Third edition. U.S. Environmental Protection Agency. Report EPA/600/4-91/002. [Edited by P.A. Lewis, D.J. Klemm, J.M. Lazorchak, T.J. Norberg-King and W.H. Peltier] 341 p. Cincinnati, Ohio.

Willemsen, A., M.A. Vaal, and D. de Zwart 1995. Microbiotests as tools for environmental monitoring. Rijksinstituut voor Volksgezondheid en Milieuhygiëne [National Inst. Public Health and Environ. Protection], The Netherlands. Rept no. 607042005, 39 p.

Zischke, J.A., J.W. Arthur, R.O. Hermanutz, S.F. Hedtke, and J.C. Helgen 1983. Effects of pentachlorophenol on invertebrates and fish in outdoor experimental channels. Aquat. Toxicol. 7: 37-58.

Appendix. Brief descriptions of the sublethal toxicity tests

This appendix gives a synoptic description of the thirteen sublethal tests, under consideration by AETE for use with mining discharges, and pre-selected for consideration here. This appendix summarizes general features of the tests in one place. The following features are covered.

- Method document
- Species and nature of the organism
- Source or organisms, and culture methods
- Life-stages tested
- Duration of exposure
- Manner of exposure (containers, etc.)
- Feeding and other care during test
- Effects measured, endpoint
- Special equipment
- Current use by other groups
- General comments (sensitivity, difficulty, etc.)

Mutatox: genotoxicity test

Document.	Manual (Microbics 1993)
Species.	Dark mutant of luminescent marine bacterium *(Vibrio fischeri*
Source, culture.	Purchase freeze-dried from Microbics. No culture, handle as a reagent
Life-stages.	Reproductive cycles during test
Duration.	24 h (Measurements at 16, 20, and 24 h)
Exposure.	Small cuvettes, purchased from Microbics
Care.	None
Endpoint.	Light production at least twice the control levels
Equipment.	Special spectrophotometer from Microbics
Current use.	Widespread
General.	Fast and low in labour. Cost low except for spectrophotometer. Sensitivity not documented in manual

SOS Chromotest: genotoxicity test

Document.	Draft instructions (EBPI 1995)
Species.	Unspecified bacterium [or bacteria?]
Source, culture.	Purchase freeze-dried from EBPI
Life-stages.	Reproductive cycles during test
Duration.	Three hours
Exposure.	Microwell plate
Care.	Overnight incubation only
Endpoint.	Optical density
Equipment.	Low-priced standard spectrophotometer
Current use.	Widespread, but not common

General.	Good reproducibility. Effect measured is unclear, in instructions

MFO in liver of trout

Document.	Preliminary report (Centre Saint-Laurent 1994)
Species.	Rainbow trout
Source, culture.	Liver cells obtained from a trout
Life-stages.	Not applicable
Duration.	48 hours for exposure
Exposure.	Four replicates of each concentration of test material plus control
Care.	Specific incubation procedures
Endpoint.	Relative quantity, messenger ribonucleic acid which controls generation of cytochrome P450IA1 (a mixed-function oxidase or MFO). NOEC
Equipment.	Fluorescent probe for RNA. Lab equipment for biochemistry
Current use.	Under development at Centre Saint-Laurent
General.	A *biomarker*, potential early-warning signal of toxic effects

Metallothionein in liver of trout

Document.	Preliminary report (Centre Saint-Laurent 1994)
Species.	Rainbow trout
Source, culture.	Liver cells obtained from a trout
Life-stages.	Not applicable
Duration.	48 hours for exposure
Exposure.	Four replicates of each concentration of test material plus control
Care.	Specific incubation procedures
Endpoint.	Relative quantity of messenger ribonucleic acid which controls the generation of metallothionein (MT). NOEC
Equipment.	Fluorescent probe for RNA. Lab equipment for biochemistry
Current use.	Under development at Centre Saint-Laurent
General.	Estimated cost for materials and time (not equipment) for both MFO and metallothionein = $125. A *biomarker*

Bacterium: Microtox chronic test

Document.	Manuals: (Microbics 1995a, b)
Species.	Luminescent marine bacterium *(Vibrio fischeri* formerly *Photobacterium)*
Source, culture.	Purchase freeze-dried from Microbics. No culture, handle as a reagent
Life-stages.	Reproductive cycles during test
Duration.	22 h
Exposure.	Small cuvettes, purchased from Microbics
Care.	None
Endpoint.	Light production as measure of numbers and metabolic activity
Equipment.	Special spectrophotometer from Microbics
Current use.	New test but will presumably become widespread
General.	Extremely fast and low in labour. Cost low except for spectrophotometer. Sensitivity not well established but claimed okay

Single-species alga: reproduction in a microplate

Document.	Methods document: (Environment Canada 1992c)
Species.	A green alga, common in N. America, *Selenastrum capricornutum*
Source, culture.	Culture repositories in Canada. Then maintain a culture in lab
Life-stages.	Complete reproductive cycles during test
Duration.	3 d
Exposure.	In wells of a 96-well microplate, 200+ μL per well
Care.	Incubation only
Endpoint.	Number of algal cells at end, compared to control
Equipment.	Particle counter or haemacytometer
Current use.	Commonly used in Canada. Similar, less convenient, elsewhere
General.	Few person-hours required. Fluorometer would be advantageous

Multi-species algae: reproduction in a microplate

Document.	Stand. operating proc.: (SRC [Saskatchewan Research Council] 1995b)
Species.	5 species of algae
Source, culture.	SRC. Reasonably straightforward but cultures must be in log phase
Life-stages.	Complete reproductive cycles during test
Duration.	2 or 3 d
Exposure.	In wells of a 96-well microplate, 200+ μL per well
Care.	Incubation only
Endpoint.	Number of cells compared to control
Equipment.	Preferably a special fluorometer for microplates, haemacytometer, shaker
Current use.	Under development
General.	Early stage of development. Using many species gives high sensitivity but culturing could be troublesome. Readings at end almost instantaneous and semi-automatic with fluorometer

Duckweed: growth

Document.	Stand. operating proc.: (SRC [Saskatchewan Research Council] 1995a)
Species.	Duckweed (*Lemna minor*) common in N. American ponds
Source, culture.	SRC preferred. Culture easy, but bacterial contamination occurs
Life-stages.	Week-old plants
Duration.	7 d
Exposure.	Petri dishes
Care.	Incubation only
Endpoint.	Number of leaves
Equipment.	None special
Current use.	Variations of test are fairly commonly used
General.	Perhaps only moderate sensitivity. Robust test. Few person-hours. No problem with turbid samples

Nematode: survival and growth

Document.	Methods paper: (Samoiloff 1990)
Species.	Small nematode (roundworm) *Panagrellus redivivus*
Source, culture.	Bioquest International Inc., Winnipeg. Culture simple
Life-stages.	New-born, grow to adults
Duration.	4 d
Exposure.	Small auto-analysis cups
Care.	None during test except temperature maintenance
Endpoint.	Survival, growth (length), maturation
Equipment.	Nothing unusual
Current use.	Not widely used
General.	Appears simple, low labour but particular handling skills. Published data-bank is small

Ceriodaphnia: *survival and reproduction*

Document.	Methods document: (Environment Canada 1992a)
Species.	A waterflea *Ceriodaphnia dubia*. N. American species
Source, culture.	Other labs. Maintain a culture in beakers or jars
Life-stages.	Neonates < 24 h old
Duration.	About 7 d, until 60% of controls produce 3 broods of young
Exposure.	Small cups or beakers
Care.	Daily feeding with mixture, new test solution daily by transfer
Endpoint.	Mortality of 1st generation, no. of young produced
Equipment.	No unusual equipment
Current use.	Widespread N. America and elsewhere. Large data-bank
General.	Generally sensitive. Some difficulties with culture, control performance

Fathead minnow: embryo-larval survival and teratogenicity

Document.	Methods document: (U.S. EPA 1994)
Species.	Fathead minnow
Source, culture.	Maintain a colony in laboratory
Life-stages.	Embryos < 36 h after fertilization of eggs
Duration.	7 d
Exposure.	Beakers or Petri dishes
Care.	Removal of dead embryos, larvae. No feeding
Endpoint.	Daily count of live larvae, versus dead and deformed larvae
Equipment.	Not extensive. Locally-made chambers.
Current use.	Widely used in U.S.A.
General.	No assessment of growth of larvae. Easier, could be less sensitive than method of Environment Canada (below).

Fathead minnow: larval survival and growth

Document.	Methods document: (Environment Canada 1992b)
Species.	Fathead minnow, small, widespread in Canada E. of Rockies
Source, culture.	Lab colonies or purchase newly-hatched larvae
Life-stages.	Larvae, hatched ≤ 24 h
Duration.	7 d
Exposure.	Small "cages" of various design in beakers or chambers
Care.	2 or 3 feedings/d, moderately time-consuming
Endpoint.	Growth, mortality
Equipment.	Not extensive. Locally-made chambers. Micro-balance
Current use.	Widespread N. America, also Europe, some method variations
General.	Care with culture (fungus etc.) and with drying and weighing larvae. Among-lab reproducibility only fair

Salmonid: survival of early life-stages

Document	Methods document: (Environment Canada 1992d)
Species.	Atlantic or coho salmon, or rainbow trout
Source, culture.	Hatcheries. No culture unless adults are held to provide gametes
Life-stages.	Newly-fertilized eggs, embryos, alevins, + fry for 30-d test
Duration.	7 or 30 days
Exposure.	Incubation cups in aquaria, renewal or flow-through
Care.	Feeding fry in 30-d test
Effects.	Mortality, abnormalities, weight of fry. Other optional items
Equipment.	Not extensive. Exposure units can be locally made
Current use.	Relatively new test. Now used in B.C., less generally across Canada
General.	Some difficulties of procedure and seasonally available gametes

Chapter 2.

Sublethal tests and their validity in predicting effects on communities in the receiving water.

General summary 40

1 Introduction 41

2 Test methods 42
  2.1 Sublethal whole-organism tests 42
    2.1.1 Faster and smaller tests 43
..... 44
  2.2 Experiments in real and artificial communities 44
    2.2.1 Larger controlled communities 44
    2.2.2 Problems with cosm experiments 45

3 Application to the Canadian mining industry of sublethal toxicity tests 46
  3.1 Implications of the review of validation studies 46
    3.1.1 Toxicity tests compared to biological surveys 47
    3.1.2 Toxicity tests compared to chemical monitoring 47
  3.2 Basic tactics for monitoring water pollution and their *raison d'être* 48
    3.2.1 Effluent monitoring: a first tactic of the triad 48
    3.2.2 Monitoring receiving-water quality: a second tactic of the triad 50
    3.2.3 Biological surveys: a third tactic of the triad 51
  3.3 Other uses of sublethal toxicity tests 52
    3.3.1 Comparisons over time or among places 53
    3.3.2 Distinguishing sources of toxicity 53
    3.3.3 Identification of toxic components within an effluent 53

4 Review of research on validating toxicity tests 55
  4.1 Validation using natural waterbodies 56
    4.1.1 U.S. EPA doses a stream in Ohio 56
    4.1.2 Eight validation studies by U.S. EPA 57
    4.1.3 Criticism of the eight EPA studies 58
    4.1.4 Re-analysis of the EPA data 59
    4.1.5 Other validations using natural waterbodies 60
  4.2 Validation using artificial stream channels 63
    4.2.1 U.S. EPA channels at Monticello 63
    4.2.2 Channels at other locations 64
    4.2.3 Channel studies of less direct interest 64
  4.3 Mesocosms in ponds 65
  4.4 Unusual experimental work of some relevance here 65
  4.5 Important reviews of data 66
  4.6 Review of the AQUAMIN interactive bibliography 69
  4.7 Literature references that were not oriented to the present review 70
  4.8 Conceptual publications 70
  4.9 Tabulation of validation studies and conclusions 71

Annotated references 73

Appendix. List of 27 reports retrieved from the AQUAMIN data-base 93

General summary

(1) The main objective of this chapter is to review the utility to the Canadian metal mining industry, of sublethal aquatic toxicity tests as tools for determining effects of effluents in receiving waters. This was done by a review of literature.

(2) A variety of rapid sublethal tests are now available, using bacteria, plants, invertebrates and fish, with test durations that are often in the range 2 days to 7 days. Some of the tests are relatively cheap to carry out.

(3) For monitoring purposes, toxicity tests on the effluent have major advantages of speed and economy over field surveys. Biological and chemical tests on an effluent complement each other by telling whether there is a problem, and if so, what it is.

(4) A complete strategy for monitoring discharges should include three components or tactics. (1) Measurements at the end of the effluent pipe, and comparison with objectives adopted for quality of the effluent. (2) Assessment of chemical and biological conditions in the receiving water, and comparison with any site-specific objectives for quality. This tactic is important for the chapter, since toxicity of the effluent can economically predict conditions in the receiving water. (3) Periodic ecological surveys to check that the first two approaches yield meaningful measurements and that goals are being met.

(5) Some attempts at validating sublethal single-species tests have made comparisons with communities of organisms in polluted rivers. That assures realism, but can entail problems of interpretation because of variable discharges or overlapping sources of pollution. Validation experiments have also used semi-natural communities in controlled or habitats, notably artificial stream channels with a diversity of biota.

(6) Twenty-nine published reports or studies were given primary attention in this review (Table 1), and varied widely in content and quality. Some comparisons were subjective, and most were general; most studies could not be called rigorous scientific "predictive" experiments. One study was rigorous and objective, and found 88% agreement in comparing the toxicity of 43 effluents with effects in the corresponding receiving waterbodies. Eight well-publicized studies in polluted rivers by the U.S. Environmental Protection Agency received criticism for inappropriate comparisons but statistical re-analysis showed a strong relationship between toxicity of receiving-water samples and instream community response. The EPA studies, like some others, had a partial program of toxicity tests on the effluent, often using instead, "ambient" toxicity tests (i.e. with the surface water). However, effluent and ambient tests appeared closely correlated.

(7) The 29 reports provided 73 useful comparisons for different locations or experiments (Table 1). Tabulation indicated 53 cases of general agreement between lab and field, and 10 cases of disagreement, for 84% agreement. Ten other cases were not directly useful because they were comparisons of values derived from the literature, because field effects were absent or inconsistent, or because they considered only lethal effects.

1 Introduction

The objective assigned by AETE that is covered in this chapter, was to conduct a literature review that dealt with the following topics.

the broad application of sublethal toxicity tests for the Canadian mining industry, and a description and evaluation of sublethal toxicity testing approaches for mining effluents

the utility of these tests as predictive/investigative tools for determining effects in the receiving water from mining effluents.

The first topic above ("broad application") is covered in sections 2 and 3. Section 2.1 gives a thumbnail history of sublethal testing and discussion of modern tests. Section 2.2 reviews briefly, the concepts involved in "ecosystem" experiments as part of validation.

Section 3 outlines the framework and rationale for monitoring discharges of effluents, and the tools that are useful in certain parts of the framework. The usefulness of toxicity tests compared to other methods of assessing water pollution is covered in section 3.1. Three tactics that should be used in monitoring are described in section 3.2. Some other common uses of sublethal toxicity tests are briefly described in section 3.3.

The second topic mentioned above (toxicity tests as predictive tools), occupies most of this chapter (section 4). This entailed a review of the literature on attempts to "validate" sublethal tests on effluents. Validation was a matter of comparing the measured toxicity with the degree of observed effect in a functioning aquatic community that received the effluent. The topic is worth serious examination; if sublethal tests on mining discharges were reasonably predictive of effects in ecosystems, their routine use could result in substantial savings on monitoring costs.

Testing of aquatic sediments is not specifically covered in this chapter since the contract was focused on tests for liquid effluents and effects of effluent. Some of the rationale and comments in section 3 apply to sediment testing, both chemical and biological, as mentioned in that section.

2 Test methods

Summary. *Sublethal tests have become faster and less laborious. The evolution has been from life-cycle tests with fish, lasting about a year, to (a) shorter tests of the most sensitive phases of the life cycle, and (b) other kinds of smaller organisms whose life-cycle exposure is 2 or 3 days or less.*

Eight tests being examined for the Canadian mining industry include organisms from bacteria to fish, and exposures from 22 hours to 30 days (see chapter 2).

Some validations have studied real polluted rivers. Controlled experiments have used many kinds of semi-natural or artificial communities. Microcosms might contain fewer than a dozen species in a container of 3 litres or less. Mesocosms range upwards in size to stream channels 500 m long, or portions of ponds or lakes. Stream channels have been relatively successful in validation, in part because a continual flow-through of water enables dosing to maintain a steady concentration, and in part because of semi-natural colonies which develop. Some ponds and enclosures in ponds have had difficulties with declining concentrations of toxicant, or poor replication of communities.

2.1 Sublethal whole-organism tests

As described elsewhere (section 3.2.3) and in the previous chapter, toxicity tests that measure responses of the whole organism are somewhat more likely to relate to the processes in an aquatic community, than are biochemical or other tests within an organism[1]. Among whole-organism responses, reproductive performance of an animal, plant, or micro-organism is often one of the more sensitive measurements of damage from a toxicant. Over the last three decades, there has been a rapid evolution of sublethal toxicity tests that measure reproduction of aquatic organisms. The evolution has been from long exposures for the entire life cycle, to short exposures focusing on susceptible stages of the life cycle. The evolution has also been from tests with fish, to tests with smaller organisms of diverse taxonomy, which have rapid life cycles and lend themselves to fast and economical laboratory tests.

Life-cycle tests are long established in mammalian toxicology, and a multi-generation test has ability to detect *any* kind of sublethal effect. For example in testing drugs for human use, the multi-generation test with white rats has been a standard and definitive examination of toxicity.

---

[1] It is risky to predict upwards through levels of integration, i.e. cells → tissues → organ systems → individuals → populations → communities → ecosystems. In essence, predicting more than two levels upwards is fruitless, and two levels is riskier than one. One looks upwards to see the significance of a change observed at a given level, and downwards to discover the cause.

Life-cycle toxicity tests for fish were pioneered as recently as three decades ago (Mount and Stephan, 1967). Initial work was with the fathead minnow *(Pimephales promelas)*, a good species for North America since it is a widespread native and takes joyfully to life in an aquarium. Males guard the eggs laid under pieces of tile, and the investigator exchanges tiles daily to count eggs and assess hatchability, without apparent chagrin among the parent fish. It requires about a year to go from newly hatched fish to the next generation. Other species such as the tropical flagfish *(Jordanella floridae)* can be tested in about three months.

McKim (1977) concluded that "life-cycle toxicity tests have provided the most reliable information on the long-term effects of toxic chemicals on fish and the basis for many of the established water-quality criteria for aquatic life".

2.1.1 Faster and smaller tests

Sublethal whole-organism tests with fish have become as fast as lethal tests, although more labour-intensive. Investigators noted that early life-stages of fish were almost always the most sensitive part of the life cycle, and month-long tests could include embryonic, larval, and early juvenile development. Results usually predicted effects in a life-cycle test within a factor of two, and predicted exactly in 83% of the cases (McKim, 1985). A recent trend uses 4- to 7-day exposures of newly-hatched fathead minnows, and takes the results as being predictive of a chronic test (Norberg and Mount, 1985).

Non-vertebrates are now tested more frequently, and favourite organisms include the small crustaceans *Daphnia* and *Ceriodaphnia* spp ("water fleas") which can go through a generation in 1-2 weeks. Algal growth or production has been tested for years, and requires only a few days for a multi-generation test of reproduction. There are standard marine equivalents in use for both fish and invertebrates.

Cairns (1988b) quotes an industrialist: "Why can't you enviromental toxicologists just give us a freeze-dried talking fish on a stick?" When the stick was inserted in the test liquid, the fish would hydrate and deliver a verbal report. Tests approaching that rapidity now exist, and they use small species including bacteria. That seems logical for protecting ecosystems because there is much greater variation in sensitivity between groups of organisms (insects compared to algae, etc.) than within a group (Klapow and Lewis, 1979; Thurston et al., 1985; Slooff et al., 1986). Therefore a variety of organisms should be tested, a principle that is recognized when deriving water quality criteria (Stephan et al., 1985). From a massive comparison, Slooff et al. (1983) suggest that screening should use not only a fish, daphnid, and alga, but also a bacterium, e.g. *Microcystis aeruginosa* which was often the most sensitive.

The trends continue today as a great variety of sublethal aquatic tests becomes available, usually making use of smaller organisms such as rotifers (Snell and Moffat 1992) for which population growth during 1.3 generations can be assessed in a two-day test with little labour.

Environment Canada has now provided a selection of modern standard methods for aquatic toxicity tests. Tests span lethal and sublethal effects, organisms from fish to bacteria, and marine and fresh water. Examples of these are provided among the eight tests being evaluated by AETE (Chapter 1).

2.2 Experiments in real and artificial communities

In the field of aquatic toxicology, attempts at field validation have involved experiments in a wide variety of biotic communities, from mixtures of micro-organisms on a glass slide to whole lakes. Many studies have used existing pollution situations in rivers (section 4).

Other studies have attempted experiments with greater control of the "field" communities, and have turned to various artificial habitats or controlled portions of natural habitats. The experimental units are variously described as "artificial ecosystems", "multispecies toxicity tests", "laboratory streams", "microcosms", "mesocosms" etc., depending on design. There are many current attempts to develop standard microcosms for testing toxicants, with a few species in a flask, carboy, or aquarium. The communities might or might not contain fish. Most procedures are in developmental stages, and there are problems in deciding what degree and kind of change represents harmful effect.

Microcosms, as the name implies, are relatively small systems, say of maximum volume 3 or 4 litres. Often they contain a few species (say a dozen or fewer) of micro-organisms including small crustaceans. Usually the whole assemblage is contained in a flask or other container in the laboratory. Replication is easy, cost is low, but because of the simplicity of the system, realism is reduced, and extrapolation to nature often remains uncertain. Variation from flask to flask can be extreme even among replicate controls.

Mesocosms might range from the size of an aquarium upwards to sections of ponds or ponds themselves. Often they are outdoors and often there is no restriction on species that are allowed to colonize. A good example of realistic mesocosms would be recirculating laboratory streams of the order of 6 m length. They can test small but complete communities, with micro-organisms, algae, invertebrates and a few fish (Warren and Davis 1971). Simple effects in the laboratory streams (biomass, diversity) are adequately predicted by single-species tests, but interactions between species are not (Hansen and Garton 1982). Assessment might include functions such as primary production. Interpretation might involve subjective assessment of complex ecosystem diagrams of energy flow and production (Warren 1971, p. 315317).

The Canadian mining industry has used such artificial streams in at least one location (British Columbia). Perrin et al. (1992) published a description of on-site, flow-through troughs as a way of assessing treated acid mine drainage. They recommended study of the algal and insect communities that developed in the "streams". The study did not carry out single-species effluent tests, but they found no effect in the streams dosed with 10% concentrations of mine discharge.

2.2.1 Larger controlled communities

A very successful approach for validation in fresh water uses a series of parallel, natural-substrate artificial streams constructed outdoors. In such a facility at Monticello, Minnesota, stream channels are of realistic size (1.4 m wide in riffle areas and widening in pools), replicates can be used, and steady flows of new water can be dosed to constant concentrations (Zischke et al. 1983a; Cairns 1985). Small populations of fish must earn their living under semi-natural conditions, and their survival, growth and reproductive success can be measured during an experiment of a few months. Other components of the community can be evaluated by standard survey techniques or by removing blocks of artificial substrate. Since an experiment involves at least half-a-dozen people for a summer, it is not cheap, but is a very meaningful approach to field validation. In at least a couple of cases, the previously-accepted water quality criterion failed to protect the stream community (Zischke et al. 1983a, Hermanutz et al. 1992).

Several groups of investigators have studied large volumes of lake or seawater enclosed in plastic bags or in "limnocorrals" running to the bottom, with toxicant added to some of the enclosures. Fish are usually excluded, and populations of other organisms are assessed by conventional sampling techniques. Enclosures are good research tools but share certain difficulties with other types of closed cosms.

At the upper extreme of manipulated communities, small lakes might be dosed and used as experimental systems, as in the Experimental Lakes Area of northwestern Ontario, where some environmental toxicology has been done by the Canadian Department of Fisheries.

2.2.2 Problems with cosm experiments

One common difficulty in studies with cosms is that constant concentrations are not maintained because toxicant is usually added only once. Thus there is lack of comparability with constant-concentration toxicity tests in the laboratory. The problem is how to approach an analysis of data based on declining concentrations of toxicant. There is seldom an attempt to maintain a constant concentration in enclosures or flask-type microcosms, although it is possible to do this by using flow-through procedures (Hedtke 1984). One exception, with documented steady concentrations of cadmium in plastic enclosures, provided valuable estimates of metal concentrations affecting copepod populations (Kuiper 1981).

Another problem with mesocosms including ponds, is that these segregated communities tend to be unstable, and replicates might diverge in characteristics for reasons that are not obvious. An example is seen in twelve ponds of 20 m on a side, constructed by Rosenzweig and Buikema (1994). Natural communities were allowed to develop in the ponds over one year. Although similar patterns of organisms developed in all ponds, the structures of the communities were never similar. After one year they still could not be used as replicated test systems. Rosenzweig and Buikema concluded that management of the communities was necessary to produce good replicates.

3 Application to the Canadian mining industry of sublethal toxicity tests

This section attempts to outline a framework of monitoring programs for water pollution from the discharge of effluents. It lists the various approaches and their purposes, and the techniques of testing and survey that fit with each. Emphasis is placed on two topics of particular concern to AETE, the use of sublethal tests on effluents, and field assessments of receiving-water communities.

3.1 Implications of the review of validation studies

Summary. *If sublethal toxicity testing of an effluent could partially replace surveys of the receiving water, there would be great advantages in speed and economy. Tests could be used frequently for good control of discharges. Biological and chemical testing of effluents supplement each other; the biology integrates all toxic influences, tells whether there is a problem, and provides an ecologically meaningful answer; the chemistry can be very fast and can tell which toxic agents are acting.*

The review in the next section (section 4) shows a strong preponderance of agreement between single-species toxicity tests and observed effects in communities or mesocosms. If sublethal effluent tests were used to monitor wastewater discharges from the mining industry, it appears that most of the time, the results would correctly correlate with environmental effects or lack of effects. The correlation, of course, would have due regard to calculations of dilution in the receiving water. The correlation of toxic effect would also have to allow for any degradation of the environment from non-toxic discharges, notably suspended solids in the case of the mining industry. If there were some initial study of the local communities and sensitive species, a suitable test (or tests) could be chosen for assessing the effluent, with improved likelihood of being environmentally correct.

An effluent toxicity test will always be a predictive test, however, with a possibility (large or small) that it will not correspond well with the actual effects in a community of organisms in the receiving water. Therefore, the field survey of the biota in the receiving water remains as the final word on whether control of toxic discharges is satisfactory. It will be necessary to carry out such surveys periodically, to check that control measures are actually satisfactory, and that the effluent monitoring is correct in its predictions. However, as information and correlations are built up for a particular location, the expensive field surveys need be done less and less frequently as a check on the routine monitoring of effluent.

Using a sublethal test for biological monitoring, instead of the old standard test of lethality, has the major advantage of increased sensitivity. As indicated in section 2.1, enormous strides have been made in supplying sensitive sublethal tests that are as rapid as lethality tests, and often as cheap or cheaper. One of the ways of accomplishing those things is to use small organisms that have a rapid life cycle and are economical to culture.

The rationale for various tactics and methods of monitoring pollution is covered in section 3.2, but certain techniques of particular interest are compared immediately below.

3.1.1 Toxicity tests compared to biological surveys

If a sublethal toxicity test on an effluent could be used as the primary means for routine monitoring of the effluent, there are some clear advantages compared to use of field surveys.

(1) *It is faster.* Results are known when the test is completed, which could be less than an hour to 4 or 7 days, depending on the test used. Complete biological surveys typically take weeks or months to carry out, analyze and report.

(2) *It is less expensive.* The cost of a single effluent test might be in the range from $75 up to about a thousand depending on the test selected. A field survey of only a control location and two downstream stations could be expected to cost $5,000 or upwards, depending on the type of habitat and the coverage of different types of organisms.

(3) *It can be used frequently*, because of the relative speed and cost. The effluent test provides some hope of detecting and correcting an unfavourable situation before it continues for very long, perhaps before any harm has been done to the environment. Frequent/rapid tests can also be associated with short-term process changes of an industrial operation, to find the most economical and favourable procedures.

3.1.2 Toxicity tests compared to chemical monitoring

There are major advantages in using a biological test (a toxicity test) to supplement chemical monitoring of the effluent. Chiefly, the advantage is in obtaining an ecologically meaningful answer; the organisms integrate all the toxic components and modifying factors, and by definition they never give a false response for toxicity. Chemical monitoring can never measure toxicity, only predict it; an organism or living system must be used to measure toxicity.

Chemical monitoring can have advantages of speed or economy, depending on the substance(s) involved. The major advantage of chemical measurements is the ability to discover *what substance(s)* is/are the likely cause of toxicity, something that a biological test cannot do. Obviously, chemical tests and biological assays supplement one another, and the two must be used together for increased understanding and effectiveness. Further comments on use of chemical procedures are given in section 3.2.2.

3.2 Basic tactics for monitoring water pollution and their *raison d'être*

Summary. *An overall strategy for monitoring should always include three tactics.*
(1) Monitoring at the end of the discharge pipe, and comparison with objectives chosen for effluent quality, usually objectives that correspond to a reasonable level of industrial practice.
(2) Monitoring of conditions in the receiving water, and comparison with site-specific objectives for quality, intended to prevent sublethal effects beyond a mixing zone. In practice, routine monitoring might be done by testing the effluent and calculating to the receiving water. This is the tactic which is being considered in the present review of sublethal testing of mining discharges.
(3) Periodic ecological surveys to check the predictions of the first two approaches, and the effectiveness of pollution control.

There is a wide spectrum of chemical, physical, and biological tests that can be used for monitoring discharges of wastewater. This spectrum can be simplified to a ***triad*** of approaches, if we consider the types of information that are obtained, and the uses for the information.

The first two tactics are: (1) monitoring of effluent quality and comparison with pre-selected objectives; and (2) monitoring of the receiving water and comparison with objectives for a quality to protect the indigenous community. Both tactics can help avoid problems before they happen, and the monitoring tests can be fast and inexpensive. However, the methods are predictive and might be in error, failing to protect the receiving community. Tactic (3), instream monitoring of the "health" of the biotic community, is the final audit of whether or not there has been deleterious effect. Surveys have no predictive capacity, however, and can only detect a pollutional effect after it has happened.

Sediment testing is not specifically mentioned, since this contract focused on tests suitable for an effluent, and validation of those tests. However, chemical tests on sediment would fit the same places in the rationale as chemical tests on surface water samples, and toxicity tests with sediment would correspond in function to toxicity tests with samples of surface water. Samples of invertebrates for a field survey would, of course, often be collected from sediments.

3.2.1 Effluent monitoring: a first tactic of the triad

Certain objectives might be adopted by an industry for the quality of an effluent at the time of discharge, without regard to dilution available in the receiving water. These objectives might be chemical or biological in nature. A common chemical example, relevant to mining, might be an objective that the effluent should not be excessively acid or alkaline, but should have pH between x.x and y.y. A common biological example that has often been adopted in Canada, would be that the effluent should not be acutely lethal to rainbow trout.

The purpose of such objectives is to achieve some reasonable standard of quality for an industry, and to avoid any severe degradation that could cause rapid damage to organisms including humans. Such objectives are clearly technology-based, so we can include under tactic no. 1, the intrinsically meaningless term "Best Available Technology" (BAT). Usually, BAT indicates application of the most recent and/or most effective waste treatment.

Since the objectives apply at the end of the discharge pipe, they have limited ecological relevance. A large-volume effluent that was non-lethal might cause sublethal effects if discharged into a small stream. Conversely, a small-volume effluent that killed trout might not cause any observable effects on organisms in a large body of receiving water.

An objective of a non-lethal effluent might tend to encourage discharge of water, and discourage water conservation in the industry. That is because the objective is conceptually based on concentration of toxic substances in the effluent, and traditionally ignores volume of discharge. We might create an hypothetical example of two industries, one discharging only pailfuls of effluent, and the other discharging thousands of cubic metres. Let us suppose that each effluent just meets an objective that has been adopted for a level of low pH, and each just meets an objective of non-lethality to fish. Hence, both effluents meet the goals, but obviously the second industry is discharging much larger amounts of acid.

A remedy that would make the objective independent of the volume of water used in the industry would be to describe all objectives under tactic no. 1, not as concentrations but as *amounts*, i.e. concentration multiplied by volume of discharge. That has been done as early as the 1970s, when objectives for Canadian pulp mills assessed the discharge of suspended solids and biochemical oxygen demand, in terms of weights. However, the weight of pollutant is customarily expressed per tonne of industrial product, i.e. a small and a large mine or mill might have the same value for discharge (kg of pollutant per tonne of product), but the large operation would obviously discharge a larger total amount. Such a system of effluent-based measurement under tactic no. 1 would still have no particular relation to protecting a given ecosystem.

If it were desired to adopt an objective for *amount of toxicity* discharged, it would be necessary to multiply the degree of toxicity (the reciprocal of the EC50, IC25, or TOEC)[2] by the volume of discharge. The result of the multiplication is usually called the *Toxicity Emission Rate*, and would be measured as "m^3 of just-toxic effluent" per unit time. It is a derivative of toxic units, which are based on the reciprocal of EC50 (or lethal concentration, for a different set of toxic units). The calculation could be based on production rate at the industrial facility, i.e. "m^3 of just-toxic effluent" per unit time, per tonne of product, usually called the *Toxicity Emission Factor.*

---

[2] EC50 = *median effective concentration*, estimated to cause a specified effect, usually sublethal, in half of the individuals in a sample of test organisms, in a specified duration of exposure. IC25 = *inhibiting concentration for a 25% effect*. It is an estimate of the concentration that would cause a designated percentage impairment in a quantitative biological function such as growth, numbers of algal cells, or luminescence of bacteria. Another number could be chosen instead of 25. TOEC = *threshold-observed-effect concentration*, the geometric mean of the LOEC (*lowest-observed-effect concentration*) and NOEC (the highest *no-observed-effect concentration)*. LOEC and NOEC are used in sublethal testing. TOEC (sometimes called in the U.S.A., the "chronic value" which can be a misnomer) can be calculated as a convenience, in order to have one number instead of two.

Persistent bioaccumulative toxicants (such as cadmium or mercury) should always be monitored under tactic no. 1. The amount discharged should be the consideration, rather than depending on site-specific objectives of tactic no. 2 (see below). Such toxicants accumulate in portions of the biosphere, and dilution should not be considered as a suitable answer to disposal.

3.2.2 Monitoring receiving-water quality: a second tactic of the triad

Monitoring the receiving water is a site-specific tactic, since concentration of the effluent depends not only on volume of discharge but also on dilution available. Clearly, *location* of an industry is of great importance, whether a tide-swept section of ocean or a slow-moving shallow creek! This is the approach of most interest for the present review. Tests on the effluent, whether chemical or toxicity tests, can be used to estimate conditions in the receiving water.

To make monitoring purposeful, objectives should be chosen for quality of the receiving water. Such water quality objectives are also site-specific because they will be chemical or biological limits that apply after dilution. Chemical objectives are usually derived from *water quality criteria* or *water quality guidelines*, scientifically derived numbers that are designed to be protective of various uses of water, e.g. "safe" for aquatic organisms (CCREM 1987).

The purpose of such objectives in the environmental field is simply to achieve a quality in the surface water that is satisfactory or favourable for aquatic organisms. There might be a mixing zone where conditions did not meet the objectives. Tactic no. 2 would often be the most useful part of a monitoring program, since tests can be carried out with relative speed and economy, and give meaningful predictions for the health of the environment.

A typical objective might be that the concentration of dissolved copper should be less than x.x mg/L for a given hardness of the receiving water. Reasonable objectives for x.x might be selected from guidelines defined in Canada (CCREM 1987), by most provinces, and widely in the U.S.A. A biological objective in standard use in the U.S.A. would be "absence of sublethal effect on (specified) aquatic organisms beyond the limits of a (designated) mixing zone".

The most efficient way to monitor for tactic no. 2 is by measurements on the effluent itself, then to *calculate* on the basis of available dilution, whether objectives will be met at the edge of a mixing zone. That is faster, cheaper and more convenient than sampling the receiving water for measurements. It also makes for easier or more sensitive measurements because the substance or quality of interest is more concentrated. To make the extrapolation from measurements on the effluent, a model of physical conditions in the waterbody is needed (currents, size, volume of flow, etc.). Another advantage of working with the effluent is the possibility of catching an unfavourable condition before it is discharged, or at least at an early stage.

As emphasized throughout this review, this is a ***predictive*** approach. It involves calculation to estimate the levels in the environment. More importantly, the chemical water quality guidelines that are available today are only the best estimates of scientists, of concentrations thought to be "safe". Similarly in the toxicity tests, what is harmful to a *Ceriodaphnia* in a test tube might not prove to be harmful to an assemblage of crustaceans in a pond. There might be changes in form of toxicant after discharge, making conditions better or worse than predicted (e.g. detoxifying of metal by binding with organic compounds, or change in degree of solubility).

3.2.3 Biological surveys: a third tactic of the triad

This third approach to monitoring studies the biota of the receiving water, and compares it with the biota in an upstream or parallel "unpolluted" location. There could also be physical and chemical surveys under tactic three (see below).

Purposes of surveys are to check that the first two tactics are making accurate predictions, and that predicted satisfactory conditions are actually being achieved. The true measurement of environmental protection is whether or not there is damage to the living communities. Findings of chemical and biological monitoring under tactics one and two are, in the end, subservient to the definitive findings of the third tactic. A survey might show that there were unexpected ill-effects, or it might show that predicted effects were nullified and did not occur.

Biological surveys are poor, however, for keeping a close record of variations in quality of a discharge. They are slow for completion of analyses (weeks or more), and can only reflect conditions that have already occurred. Since the surveys are usually infrequent, there is little opportunity for detecting problems at an early stage and curing them.

It is not necessary to assess the entire assemblage of organisms in a community, which would require many taxonomic experts at some expense. Surveys need not be huge, and can focus on a segment of the community. In fact there is great redundancy of information provided by the different components of an ecosystem, i.e. it would not be unusual to have similar indications of damage among the fish, among the aquatic insects, the molluscs, the periphyton, etc. For example in freshwater streams, conclusions based only on mayflies and beetles were essentially the same as from a massive survey of biota ranging from diatoms to fish (Kaesler et al. 1974).

The best indicators of pollutional status are organisms that: (a) remain fixed in one limited area in the receiving environment; (b) have a life cycle of intermediate length, of about a year; (c) show a range of sensitivities to pollutants among their various species; and (d) are relatively easy to sample and identify. The macroinvertebrates living on the bottom of the waterbody are generally conceded to best fit those requirements. Fish are usually poor indicators because they might have moved around too much, in and out of a polluted zone. Micro-organisms have the disadvantage of a fast life cycle, so that their populations might recover from a recent event of pollution and fail to give evidence of it. Properly designed surveys can be infrequent since they are only checks of the more immediate control measures, and since survey results are retrospective for weeks or months (using macroinvertebrates).

Biochemical/physiological testing of resident fish or other organisms might supplement the ecological survey. Such use of "biomarkers" is currently receiving much attention and energy. However it is extremely risky to substitute this approach for part of the ecological survey. Use of within-organism variables would involve, again, prediction upwards over several levels of integration to community effects, with possible erroneous conclusions that would defeat the purpose of tactic number three. Use for this purpose would require verification that an effect was directly associated with meaningful changes of the whole organism (growth, reproduction etc.) or of populations and communities. Physiological measurements might well explain *why* some effect was observed at the population level. Biomarkers might also serve as a convenient early warning signal of sublethal effects. That has been elegantly shown by a series of within-fish derangements near Canadian pulp mills (Munkittrick et al. 1994).

Physical and chemical surveys could be important under tactic three, for such purposes as defining the plume and concentration gradients, or decay of pollutants downstream[3]. Another role would be to identify the toxicants responsible for any observed biological effects. Chemistry should definitely be used if there were potential discharge of a persistent bioaccumulative toxicant; indications of deleterious accumulation might be detected much earlier with chemistry than with biological methods.

All three tactics in the triad have a role in monitoring discharges. The one of primary interest in this review is tactic two, since it includes sublethal tests on an effluent to predict the degree of effect (if any) in the receiving water.

3.3 Other uses of sublethal toxicity tests

Summary. *Some other uses of sublethal tests can be mentioned.*
(1) Documenting improvements or other changes in wastewater quality over time.
(2) Distinguishing the more important sources of toxicity, among several waste streams within a given industrial operation.
(3) Distinguishing the substances within a waste stream that are the chief causes of toxicity.

A primary focus of this chapter is the validity of toxicity tests for predicting effects in the real world. That question of valid prediction is a major one for their use in monitoring Canadian metal-mining effluents. There are other functions of toxicity tests, however, and they are briefly mentioned here.

The sensitivity of sublethal tests makes them more useful than lethal tests for all purposes. If a lethal test does not produce an effect in full-strength effluent, a common finding nowadays, then one has no measure of the degree of toxicity of the effluent. Sublethal tests can be expected to be an order of magnitude more sensitive, and are thus more likely to produce a useful quantitative measurement. The speed of modern sublethal tests makes it feasible to use them for many purposes. Most sublethal tests use exposures of 2 to 7 days, similar to the customary exposures in lethal tests with aquatic organisms. Bacterial tests might be faster, for example a true chronic Microtox test uses a one-day exposure.

---

[3] Most surveys of chemical conditions would analyze their results in terms of water quality objectives, i.e. they would be part of tactic number two. Chemical surveys of toxicants (say copper) in the receiving water (or sediments) can never provide proof of damage under tactic no. 3, they can only predict it under tactic 2, when calculated concentrations are estimated to exceed quality objectives.

3.3.1 Comparisons over time or among places

The basic questions answered by a toxicity test are "Is it toxic?" and "How toxic?". The second question is answered by the quantitative endpoint of the test. For a metal-mining wastewater, the endpoint would be in terms of percent concentration of the waste, as the IC25 or NOEC.

The IC25 or other endpoint could be used in many types of comparisons. A useful one might be in demonstrating the changes in toxicity over time, presumably a decrease over time. Or it might be desired to compare the toxicity of wastewater from discharges at different mines, or among different categories of mines (base metal compared to gold, etc.). All such comparisons can be made, given a definitive endpoint from a standard toxicity test.

3.3.2 Distinguishing sources of toxicity

If one were engaged in reducing the toxicity of wastewater discharge, it would be appropriate to focus on the major sources or causes of toxicity. Most industrial discharges are composed of a number of waste streams with different origins around the site, or different sources in the process of milling or manufacturing. One can use toxicity tests to rate the different sources or component streams of waste. One could also use chemical measurements if the important toxicants were known, or better still a judicious combination of chemical and biological testing.

A classic example of toxicity sleuthing within one industrial operation was for a pulp mill in northern Ontario (Scroggins 1986). Toxicity tests were carried out on the wastewater in the final (combined) sewer, and also for the several component sewers from various parts of the mill, which contributed to the final effluent. Toxicity was expressed as "Toxicity Emission Factors" (TEF, see section 3.2.1). The processes (and sewers) which were the most important contributors of toxicity were easily identified by the magnitude of the TEFs. When two sewers combined, the TEF for the combined waste was either similar to the sum of the TEFs for the individual sewers, or else was smaller. That indicated generally additive to less-than-additive toxic effects when the streams combined. The usual finding in the pulp mill was moderate less-than-additivity. For example, when 10 streams from one section of the mill combined, the arithmetic sum of their individual toxicities (as TEF) was 980, but the measured toxicity was only 730 TEF. In other words, the net effect was as if 74% of the toxicity in individual streams added together and was retained in the combined wastewater. From this work, Scroggins (1986) was able to identify the more toxic waste streams, i.e. the parts of the mill where it would be logical to concentrate efforts for effective reduction of overall toxicity.

A similar approach could use chemical testing of the individual wastewater streams to assess the major sources of toxic substances, if they were known. Using both chemical and biological tests would, as usual, be a more powerful approach.

3.3.3 Identification of toxic components within an effluent

A common procedure nowadays for improving the quality of effluents is the "toxicity identification evaluation" *(TIE)*. Beyond the clumsy name is a set of pragmatic techniques which are very useful in identifying the substances within an effluent which cause its toxicity (Norberg-King et al. 1991). For example, the techniques were used on a well-treated wastewater at a South Dakota mine, in an attempt to identify residual toxicity; unfortunately the techniques were not successful in that case (Times Ltd. 1994).

TIE involves a combination of chemical and biological procedures which are designed to identify certain classes of toxicants in the wastewater, or else eliminate a class as a factor. A particular physical or chemical manipulation is carried out on the sample of effluent, then its toxicity is measured to assess whether the original toxicity remains or is decreased. The technique is repeated with other manipulations until the main categories of toxicants are identified. Manipulations include aeration, filtration, extraction, chelation, oxidant reduction and/or complexation with sodium thiosulphate, and change of pH. After each manipulation of an aliquot of wastewater, a test is done to see if toxicity has changed, and depending on the result, inferences can be made about the class of substance that was causing the toxicity. An example relevant to mining wastes would be a decrease in toxicity after adding the chelator EDTA, an obvious indication that metals were contributing to the toxicity. Similarly, oxidative compounds like chlorine would be expected to have their toxicity reduced by the addition of sodium thiosulphate. If ammonia were a major contributor, toxicity should be lower at pH 6.5 than at pH 8, although many metals also change their toxicity with pH and might complicate the picture.

The combination of toxicity testing and physico-chemical procedures is fruitful in this aspect of pollution control as in many other aspects.

4 Review of research on validating toxicity tests

Summary. *Eight validations in polluted rivers by the U.S. Environmental Protection Agency had acceptable tests and surveys, but were criticized for inappropriate methods of comparison. A re-analysis by robust canonical correlation analysis showed that the relationship between toxicity of receiving water samples and instream community response was strong, and in fact showed 90% agreement. Like many of the studies, this one did not carry out a strong program of effluent tests. It appeared that there were consistent relationships between toxicity tests in the effluent and in the receiving water, when both were done.*

Twenty-one additional published studies were included as primary material for this review (Table 1), and varied widely in content and quality. Some comparisons were subjective and most were general, and could not be considered to be rigorous scientific "predictive" experiments. One study was rigorous and objective, and found 88% agreement in 43 comparisons of effluent toxicity with effects in the receiving waterbody. Two studies were major literature reviews; both concluded good relationships between laboratory tests and ecosystem effects, or single-species to multi-species assessments.

A tabulation of the individual comparisons indicated 53 cases of agreement between lab and field, and 10 cases of disagreement, an 84% rate of agreement. Ten other cases were excluded from that tabulation because they were inconsistent, had no effects in the field study, compared lethality not sublethal effects, or were literature searches without direct experimental comparisons. That tabulation is dominated by the one publication that compared 43 locations; if each publication were counted only once, the rate of agreement would be 74%.

Purpose. This section attempts to gather what information is available in the literature, about whether toxicity tests on an effluent can be used, with due regard to dilution, to predict the degree of toxic effects in the community of aquatic organisms receiving the effluent.

As described in section 2.2.5 of chapter 1, it is always risky from an ecological point of view, to use knowledge gained at one level of biological integration, to predict what might happen at a higher or broader level of integration. In aquatic toxicology, reproductive success of individual waterfleas in the lab does not necessarily predict what will happen to populations of crustaceans in the wild, and effects on a species are not likely to predict exactly, the changes in a community. Accordingly it is highly desirable to assess the predictive value of single-species laboratory tests by comparative studies of effects in a functioning community. In this chapter and report, such assessment is termed ***validation*** without necessarily implying that there will be positive confirmation.

There has been appreciable research on validation, but much of it fails to make formal comparison of (a) effects in a single-species effluent test with (b) well-being of an aquatic community. Most of the satisfactory comparisons use natural communities in a polluted river, while some use artificial communities or cosms with varying degrees of realism (section 3.2). All kinds of comparisons are reviewed here.

4.1 Validation using natural waterbodies

Much of the general information available, unfortunately does not provide specific comparisons that are useful for our purposes. To some degree, the thousands of local biological surveys of pollution provide comparisons. The biggest problem in most of the survey results is that there is no clear comparison with any single-species effluent test -- that was not part of the work. The bibliographic data-base of AQUAMIN contains 743 references to environmental studies at metal-mining sites in Canada (mostly), but unfortunately it did not provide formal comparisons which could be used in this review (see section 4.6).

Nevertheless, field studies give some general confirmation of laboratory estimates of harmful concentrations. This may be seen in documents used to develop water quality criteria by EIFAC, the European Inland Fisheries Advisory Commission (Alabaster and LLoyd 1980). For various water quality characteristics, EIFAC reviewed the literature on*both* laboratory and field work, and most results fitted together reasonably well. From the combined data from field and lab, EIFAC generated water quality criteria for 10 pollutants or environmental conditions[4].

There can be complications in using a real waterbody as a base of comparison. Common problems would be that (a) there was more than one active pollutant or effluent affecting a region, and (b) concentrations would usually vary with time and might show extreme fluctuations.

4.1.1 U.S. EPA doses a stream in Ohio

An ultimate validation experiment involved treating a small river to a constant concentration of copper for two years, and assessing wild and confined fish and other biota. This was done in Ohio by U.S. EPA (Geckler et al. 1976). We may never see another test on that scale because of the expense of instrumentation and dosing. EPA did life-cycle tests with fathead minnows in a streamside lab. The lab tests on toxicity of copper yielded good agreement with population studies of invertebrates, and with performance of fish held in cages in the stream[5].

---

[4] Suspended solids, extreme pH, temperature, dissolved oxygen, ammonia, phenols, chlorine, zinc, copper and cadmium.

[5] The agreement occurred when dilution water in lab tests was taken from upstream in the river, because of detoxifying characteristics of the water. That is normal practice in such situations.

There was, however, a major surprise with wild fish in the creek. Avoidance reactions by fish were most important, and occurred at concentrations lower than those causing physiological effects. Wild fish moved downstream, out of some sections of stream where conditions were physiologically suitable for them, as determined by spawning performance of caged fish. When the wild fish reached the downstream barrier, they could go no further, and they too spawned. An avoidance response had not been studied in the Ohio streamside laboratory, nor predicted.

Accordingly, this experiment provided two major conclusions. (1) The lab tests agreed with, and predicted, the physiological effects which prevailed in the stream at given concentrations of copper. (2) The lab tests did not predict the total ecosystem effect because of the avoidance reactions. The second conclusion emphasizes the need for periodic field surveys in the receiving water, to check that predictions from monitoring the effluent are correct.

4.1.2 Eight validation studies by U.S. EPA

In the 1980s there was a deliberate attempt by U.S. EPA, spearheaded by Dr. Donald I. Mount, to test whether the 7-day sublethal tests with larval fathead minnows and *Ceriodaphnia* correlated well with observed conditions in receiving waters. The program was an extensive one at selected U.S. sites thought to be degraded by one or many effluents.

The eight studies differed somewhat in details, but usually there were single-species toxicity tests on effluent and on samples of receiving water ("ambient tests"). Simultaneously, there were standard field surveys of water chemistry, resident fish, macroinvertebrates, and usually plankton and/or periphyton. Hydrographic work identified plume concentrations at discharges. The work produced some surprising findings of reduced toxicity of combined pollutants, but reasonably good agreement of single-species tests with the effects on resident communities.

> *Mount and Norberg-King (1985).* A small creek in an agricultural area received discharge from a chemical plant, but it proved to be non-toxic in the single-species tests. Similarly, there was no biological effect in the stream.
>
> *Mount and Norberg-King (1986).* There were numerous discharges into 125 km of the Kanawha River in West Virginia. This study did not plan to test effluents for toxicity; a few such tests were done, however they failed to predict toxicity found in the river. Ambient sublethal tests with *Ceriodaphnia* showed good correlation with numbers of zooplankton species, but underestimated the effects on macroinvertebrates.
>
> *Mount et al. (1984).* The Ottawa River in Ohio received effluent from a municipal treatment plant and a refinery. The authors studied toxicity of the effluent with both sublethal tests; ambient tests used *Ceriodaphnia*. Instream effects ended where the ambient toxicity tests indicated no effect, and the authors considered that effluent and ambient toxicity tests predicted conditions in the receiving water accurately. This study included extra features such as lethality tests, caged fish, and insect drift in the river.

Mount et al. (1985). On this river in Alabama, effluent toxicity predicted downstream effects of three coke plants and a municipal treatment plant. Numbers of species below the discharges were indeed reduced by one half or more. At many other stations, there was correct prediction of no deleterious effect. No single test species or community group was suitable for assessing impact at every station.

Mount et al. (1986a). In the Naugatuck River of Connecticut, toxicity tests with effluent were not available. Ambient toxicity tests were compared with damage shown by instream surveys, with a claim of good correlation.

Mount et al. (1986b). In this estuary near Baltimore, it turned out that there were too few species for an adequate survey of pollutional changes. The freshwater toxicity tests were run with parallel tests to watch for salinity effects. The predictions from tests with effluent agreed with results from tests with surface water from the estuary.

Mount et al. (1986c). Many effluents were discharged into the 12-km study section of the Ohio River in West Virginia. Here again, no toxicity tests were done on the effluents. Variation in sublethal ambient tests with *Ceriodaphnia* showed general parallelism with the number of species of resident macroinvertebrates.

Norberg-King and Mount (1986). A stream in agricultural Oklahoma was affected by an oil refinery, fertilizer plant and municipal plant. Predictions from effluent tests identified the most affected station; the authors claimed excellent agreement between lab and field.

It is not profitable to examine the original conclusions from these 8 studies because of certain shortcomings, but the overall findings are discussed below. The toxicity tests and survey work were satisfactory in most cases, but the analyses of results have been severely criticized. Certainly some very individualized methods were used in the comparisons. For example the degrees of effect were often measured as percent change from a "normalized" value, which was established as the best performance among all sampling stations; customarily the comparison should be made with an "upstream" control, or a clean location on a nearby similar waterbody.

4.1.3 *Criticism of the eight EPA studies*

Criticism of the general approach was provided by Cairns et al. (1988). They pointed out that the EPA studies established a simple correlation between results of the single-species lab tests and the instream effects, but such a correlation did not necessarily mean that one was a dependable predictor of the other. The work was suitable for a preliminary study, i.e. a correlation. To actually validate the predictive power of the lab tests, the next stage should have been to return to the sites, test the effluent toxicity, predict the range of instream effects, then measure them and assess the goodness of prediction. From a formal scientific point of view, Cairns and colleagues are (see the annotated reference list for some of their words). However, in view of the large amount of pre-existing work on toxicity and field surveys that serves as a background, most workers would no doubt be convinced by a good initial correlation.

Rather devastating criticism of the EPA methods of analyzing the results was levelled by Marcus and McDonald (1992). They politely say that the EPA comparisons were "mathematically inappropriate", and that the design of the studies meant that inferences from the results could not be applied to other places or times. However, Marcus and McDonald carried out a partial evaluation of the EPA results using canonical correlation analysis, and their conclusions are encouraging. They found reasonable relationship between the variables in sets of laboratory toxicity and field data, for most of the studies. They wrote "that the short-term chronic *Ceriodaphnia* and 7-d fathead minnow tests used to determine ambient toxicity can provide useful information about biological community structure in stream waters where ambient toxicity has a controlling influence".

From these criticisms, it appears that the eight EPA validation studies were satisfactory in technical parts of surveys and testing. They showed correlations between single-species tests and communities, but design and analysis meant they could not really establish predictive relationships that could be applied elsewhere.

4.1.4 *Re-analysis of the EPA data*

Apparently the U.S. EPA became cognizant of the limitations of their analyses of the eight studies, and commissioned a group of scientists to statistically re-examine the results. They called their analysis a "robust canonical correlation analysis" (Dickson et al. 1992).

Dickson et al. found that the most useful variables in the toxicity tests were neonate production by *Ceriodaphnia* and dry weight of fathead minnow larvae; most useful in the river surveys were numbers of species of fish and of invertebrates. Their analyses "demonstrated that statistically significant relationships between ambient toxicity and instream impact existed", and that "the relationship between ambient toxicity and instream biological response is strong".

A pie diagram of one of their results (Fig. 1) shows that overall there was 90% agreement of lab and field work. This is the same information presented in tabular form in the annotated reference list under Dickson et al. (1992). It is derived from a contingency table with 94-95 percentile cutoffs applied to the data in a binomial model, in order to minimize the chance of erroneously calling a polluted site clean (Type 2 error of statisticians). Similar agreement was obtained if other cutoffs were applied, for example, to minimize the chance of calling a clean site polluted.

We may accept the conclusions of Marcus and McDonald (1992) and Dickson et al. (1992) that the EPA studies showed reasonable correlation between sublethal toxicity tests and the effects in field surveys. Apparently, the same conclusions about effect of a discharge would be drawn most of time, whether the approach was toxicity testing in the lab or surveying in the field.

Correlations from the EPA work were mainly based on "ambient" toxicity tests, i.e. done with samples of surface water, not effluent. That should not be a great handicap since there would usually be an uncomplicated relationship between the results of effluent and ambient tests; it would be largely a matter of calculating the dilution involved, which could be estimated by physical methods with relatively little error [6]. Therefore the general statements about correlations of ambient tests could be taken to apply to effluent tests, with minor reservations. Most investigations find agreement, e.g. *in situ* tests with salmonids "agree closely with laboratory-derived toxicity findings" for zinc, copper and cadmium (Davies and Woodling 1980).

---

[6] Some caution should be exercised. The effluent test should use dilution water taken from the same body of water, upstream of the effluent. Modifying factors would be taken care of in that way, and they can have major effects on toxicity (Sprague 1985, Persoone et al. 1989). For example any detoxifying substances in the surface water would be included, as would any added effects of upstream toxicants. The toxicant might degrade more readily in the natural environment than in the laboratory.

Figure 1. Results of canonical correlation analysis of the amalgamated data from the eight validation studies of U.S. EPA, as re-analyzed by Dickson et al. (1992).

The diagram represents 80 pairs of lab/field data and is for 94-95% cutoff of the data, to minimize Type 2 error. The biggest portion of the diagram represents an impact predicted by the lab tests and observed in the field survey; that case and the opposite case of no impact represent 90% agreement of the lab and field work. Diagram from U.S. EPA (1991).

4.1.5 Other validations using natural waterbodies

A Kentucky stream yielded encouraging correlation in a high-quality study by Birge et al. (1989). Lethality and teratogenicity in 8-day tests with fathead minnow larvae were studied; lethality was more sensitive, so (paradoxically) it become the criterion in the sublethal test. Exposures of larvae estimated that negligible mortality (i.e. a threshold of effect) would occur at 36% effluent. The field work agreed that the station with no significant effect (fish or invertebrates) had a concentration of 33%. An upstream station with 53% effluent did not show an effect on fish, but there were only 22 species of invertebrates compared to control values of 30 and 34, and diversity index of only 2.5 compared to the control's clean-water indices of 3.3 and 3.7.

The ambient toxicity tests showed the station with 33% effluent to be the first one that did not cause detectable toxicity of the surface water. The agreement of the effluent toxicity test with the ambient one is to be expected. The prediction of a 36% threshold from the effluent tests must be considered good agreement with the field finding of 33%. Mortality in the ambient toxicity tests showed good correlations with the field surveys of macroinvertebrates (0.93 with diversity index and 0.96 with number of species).

In North Carolina, 43 effluents and their receiving waters were assessed in an objective manner by Eagleson et al. (1990). Toxicity was assessed by 7-day tests on the effluent with *Ceriodaphnia*, then effects were *predicted*, according to whether the calculated "safe" concentration would be exceeded at low flow. Observed effects in the waterbody were decided by statistically significant changes in macroinvertebrates. The overall agreement of the predictions with the field observations was 88%, distributed as in Figure 2.

Figure 2. Agreement of objective predictions from toxicity tests on effluents, with the observed impact in receiving-water communities, for 43 discharges in North Carolina (Eagleson et al. 1990). Diagram from U.S. EPA (1991).

Researchers in Virginia assessed a single effluent with several techniques (Pontasch et al. 1989). Of relevance here are the results for reproduction of *Ceriodaphnia* showing 1.7% effluent as the TOEC (threshold-observed-effect concentration). That agreed well with a threshold of effect of 2.0% for field surveys of macroinvertebrates[7].

---

[7] Of subsidiary interest, a microcosm study in the field (protozoan colonies exposed in the stream) yielded a TOEC that was 4- or 5-fold higher, i.e. the microcosm was more tolerant. However, microcosms exposed in the laboratory were 5- to 6-fold lower, i.e. more sensitive than the field survey and *Ceriodaphnia*.

A metal mine in South Dakota had an effluent that consistently showed no effect in sublethal tests with fathead minnows, but was usually toxic in reproductive tests with *Ceriodaphnia* (Times Ltd. 1994). There were 20 tests with both species during two recent years. A single survey of benthic invertebrates 1.6 km downstream of the minesite showed no effect compared to the upstream control [8]. This case is included as a disagreement in the comparisons here.

A Colorado stream showed fewer resident salmonid fish, at concentrations of metals from mining that should have been "safe" according to pre-existing data from sublethal testing in the laboratory (Davies and Woodling 1980). *In situ* tests with fish agreed with the field findings.

In a New Brunswick river, migrating salmon appeared to show avoidance reactions when copper and zinc from mining pollution reached 0.4 toxic units (40% of the lethal level expected from laboratory work, see annotation to Lloyd and Jordan 1964 in the reference list). Laboratory work with fingerling salmon showed avoidance reactions at 0.02 toxic units, lower than the in-stream finding by a factor of 20. In this case the effects in the field were much less than in the laboratory. Avoidance behaviour is, of course, a special subject and the poor prediction was almost certainly because of strong motivation in adult fish on a spawning migration, compared to lack of preference by fingerling fish for one end or the other of their plastic trough in the laboratory. This finding is relevant to mining pollution, but cannot be used in this review since it is not in the same category as the comparisons of whole-organism damage.

Some results of less direct usefulness for our present purposes, are provided by other studies that used communities in natural waterbodies. In Colorado, an assessment of streams affected by metals from mining wastes was inconclusive because of contradictory findings in both the laboratory and the field (Clements and Kiffney 1994). Reproduction of*Ceriodaphnia* appeared to be affected by water from the reference station and downstream effects were contradictory. Nor was there a clear pattern of degradation and recovery of the macrobenthic community in the river downstream from the source of pollution. This study cannot be regarded as either a correlation of lab/field data, or a contradiction between lab and field.

Laboratory tests on an "anti-pollutant" for zinc and copper were verified by a test in a small stream. In a section of stream dosed with high concentrations of the metals, fish died within 5 hours. Downstream of that, where the anti-pollutant chelator was added, fish survived the four-day exposure of up to 6 times the normally-lethal level of metals, with no apparent disturbance of behaviour (Sprague 1968).

---

[8] Although the report by Times Ltd. (1994) says it includes the "results of case studies demonstrating no meaningful correlation between [effluent] tests and instream water quality impacts", only one assessment of the downstream community was found. The report describes unsuccessful efforts to identify the cause of toxicity for *Ceriodaphnia*. The report launches a strong attack on toxicity tests with *Ceriodaphnia*, because of the variability of tests, variation in natural reproductive performance, and using a non-native species. It does not attack the sublethal tests with fathead minnows, which usually showed no effect and passed regulatory requirements of the U.S. EPA.

Robinson et al. 1994 studied 11 Canadian pulp and paper mills. They carried out sublethal tests with *Ceriodaphnia* and fathead minnow larvae on the receiving water upstream and downstream of the mills, although not on the effluents. They did not do conventional community studies but made non-statistical comparisons with pre-existing information on such studies. They concluded that the sublethal toxicity tests were generally correlated with historical data on benthic macroinvertebrate community responses. (Physiological disturbances in fish were found at lower concentrations, but it was not shown whether, or how, those disturbances would relate to population effects.)

4.2 Validation using artificial stream channels

The best validation in this category would be sublethal tests in the laboratory, with prediction to community effects in large, semi-natural controlled and dosed streams in replicated exposures. This section arranges the various studies in two sections which approach that ideal, then another section on studies that are of less direct usefulness for the present review.

4.2.1 U.S. EPA channels at Monticello

This research station in Minnesota has eight artificial channels, each 520 m long, and customarily used with flows of 0.76 m^3/s. Natural water from the adjacent Mississippi river flows through them. A number of toxic substances have been tested in the channels, usually with replicates.

A good example of the use of artificial outdoor streams is work on pentachlorophenol by Zischke et al. (1983a). Laboratory tests estimated a "safe" concentration of 48μg/L, and it was close to being correct as determined by stream tests. At that concentration fish proved more sensitive than macroinvertebrates and other organisms. For fish, there was a small reduction in growth and increase in drift of larval fish, although no effect on reproduction. Exposures lasted 12 weeks; there were no replicates although tests were repeated in a second year.

Tests with selenium showed the importance of mesocosm validation for some toxic substances (Hermanutz et al. 1992). Historic laboratory tests had estimated 26μg/L of selenium as a "safe" level, but a one-year exposure of fish in artificial streams showed that even 10μg/L had appreciable sublethal effects on fish (growth, survival of young, internal damage). In the channels, intake via the food had increased toxicity, compared to the previous laboratory tests in which the surrounding water was the only route of exposure.

For *p*-cresol, only acute lethality was estimated in the laboratory, with thresholds in the vicinity of 10 to 20 mg/L for three species of fish, and 2 to 5 mg/L for the crustaceans*Daphnia magna* and *Hyalella azteca*, and a damselfly. Artificial channels were dosed with 8 mg/L for only one to four days. Effects on species, numbers, biomass, and "community" variables were reasonably predicted by the single-species tests. Survivorship rates were consistent, and although community effects were indirect through metabolism of aquatic plants, the correspondence of concentration/effect was retained. This comparison at lethal levels is not useful here.

Some of the studies at Monticello did not do specific laboratory work for direct comparison, in part because there was already an extensive body of information on toxicity of the substance. That was the case for work on the insecticide diazinon (Arthur et al. 1983) and acidification (Zischke et al. 1983b), which are mentioned here as good examples of research in artificial streams. In those studies, diverse findings on species sensitivity were products of the research, as well as effects on communities.

4.2.2 Channels at other locations

Some excellent British work tested a mixture of chlorinated ethers from a petrochemical plant (Crossland and Mitchell, 1992). The sublethal lab tests showed a single-species NOEC of 1.0 mg/L for growth of *Daphnia magna*. Measured constant concentrations in artificial streams assessed invertebrates, with NOEC = 0.44 mg/L for feeding of*Gammarus pulex*. This must be considered reasonable agreement (2.4-fold difference).

Sublethal and sub-acute laboratory tests with fathead minnows and the amphipod *Hyalella azteca* showed TOECs from 0.42 to 1 mg/L of the detergent LAS (Fairchild et al. 1993). Experimental streams in Missouri (50 m long) were run just lower than that (0.36 mg/L), and showed no effect on periphyton, macrobenthos, or biological processing of dead leaves during 45 days. This indicates agreement, so far as the stream test went. This study does not, however, provide a useful comparison because it did not show whether dosing the stream above the TOECs (laboratory) would have caused an effect in the streams.

In Oregon, 5 sublethal single-species lab tests of mortality, growth and fecundity "adequately predicted the concentrations of diflubenzuron which affected ... stream communities" (Hansen and Garton 1982). They measured biomass and diversity in artificial streams housed in a laboratory, with communities established for 3 months and then dosed for 5 months.

A large 5-laboratory effort co-ordinated by EEC (1992) appears to have found agreement between laboratory and mesocosm for 3 chemicals (copper, lindane and atrazine). For the fourth chemical (3,4-dichloroaniline) there were community effects at about one-tenth of the lab NOEC. These conclusions on correlation are adopted as valid for the present review, although the report is fragmented and detailed to the point that an objective rating is difficult. The participating laboratories developed new sublethal tests with two planktonic algae, two protozoans, a rotifer, and other macroinvertebrates. Experiments in artificial streams and in pond enclosures lasted for one and two months.

4.2.3 Channel studies of less direct interest

In British Columbia, treated acid mine drainage was assessed in on-site, flow-through troughs (Perrin et al. 1992). There was no effect of 10% concentration. Since the use of mesocosms for a mine discharge is certainly relevant to the present review, it is unfortunate that the study did not have single-species effluent tests to provide a useful comparison.

A study in Virginia showed that the communities in artificial streams behaved similarly to those in real streams, when dosed/polluted with copper. No single-species lab tests were used, so the work is only of indirect usefulness for the present review.

4.3 Mesocosms in ponds

Various approaches have been used to carry out validation experiments in ponds. A replicated series of small ponds is a logical technique, but natural colonization of such ponds is more variable than might be thought (see Rosenzweig and Buikema 1994 in reference list). Closing off a section of a natural or artificial pond is a useful technique which can allow good control and easier replication.

Perhaps the best correspondence between lab and cosm was obtained by Larsen et al. (1986), who tested the pesticide atrazine against 8 species of algae. Excellent agreement was reported, with 50% inhibition of photosynthesis, respiration, and algal biomass occurring at concentrations of 100 to 155 μg/L in single-species tests, in microcosms, and in experimental ponds.

Communities in enclosures in a pond in Ohio were more sensitive to copper than found in 7-day sublethal lab tests with *Ceriodaphnia* (Moore and Winner 1989). Although the well-being of *Daphnia* and macroinvertebrates in enclosures were correctly predicted at the "safe" concentration of copper, there was a decline in populations of algae, rotifers and some copepods.

4.4 Unusual experimental work of some relevance here

The two studies mentioned here are from classic work of the British Water Pollution Research Laboratory in the 1960s. They do not fall in with most of the other work reviewed because they deal with lethality rather than sublethal effects, but they give some affirmative support to the prediction of toxic effects from laboratory to the field.

Figure 3 from Herbert (1965) shows predictions from laboratory lethal tests, to a river that was apparently heavily polluted with discharge from a gas plant. The major toxicants in the effluent were measured chemically, and their effect predicted from known lethality curves developed in the laboratory. The effect of mixtures was integrated by the*toxic units* method (see annotation for Lloyd and Jordan 1964, in the reference list). The mortality of caged fish was followed daily in the river, with replacement of fish. The degree of mortality is shown by the black circles in Fig. 3. Herbert indicates that predictions were 83% correct.

Lloyd and Jordan (1964) studied 24 effluents for their direct toxicity, and used chemical measurements of the chief contaminants to calculate toxic units. They indicate that the toxic units method gave a correct assessment of the lethality of 19 effluents (Fig. 4), which is 79% agreement.

Figure 3. Observed and expected mortality of caged fish in a British river polluted by gas-plant waste. From Herbert (1965).
The observed daily mortality is shown by the proportion of the circles that is black. When the sum of toxicity of chemical constituents of the water rises above the horizontal line (1.0 toxic unit), mortality is expected.

4.5 Important reviews of data

Two published reviews of "field validation" did a remarkably thorough examinations of the experimental literature. Conclusions of those scholarly reviews are accepted here, although the findings are not used in the tabulation of agreement (sections 4.9 and 4.10). Both reviews concluded that good predictions can be made from single-species tests to multi-species assemblages. Both reviews were published by people from the Netherlands, most of them employees of the National Institute of Public Health and Environmental Protection. Whether their affiliation lends credence to the conclusions is left open to the reader, but excellent interpretive work in environmental toxicology has come out of the Netherlands over the last decade.

Figure 4. Observed and expected mortality of fish in 24 British effluents. From Lloyd and Jordan (1964).
A toxicity index of 1.0 is expected to be just lethal. Observed values are from toxicity tests with fish in the effluents. "Predicted" toxicity index was calculated from chemical concentrations of the constituents of the effluents, and their known toxicity, with components summed.

The masterful review by Slooff et al. (1986) correlated single-species and multi-species tests. Although it dealt with acute toxicity for the most part, the sublethal parts of the conclusions are of relevance here. In particular, the review concluded that multi-species or "ecosystem testing does not lead to results that are dramatically different from those obtained with single-species tests". From the data they considered reliable, the authors calculated a predictive relationship as follows.

$$\mathrm{NOEC(ecosystems)} = 0.63 + 0.85 \log [\mathrm{NOEC(single\text{-}species)}]$$

For this, $r = 0.85$.

That relationship deals with sublethal effects. Accordingly, the conclusion is that single-species effluent tests can indicate effects in receiving-water communities with reasonable accuracy.

Another conclusion by Slooff et al. was that chronic toxicity can be predicted with fair reliability from acute toxicity; they gave a formula for this (see reference list). Another major conclusion was that there was great individual variation among tolerance by different species, whether they were from the same general taxonomic group or different ones.

Another massive review of the literature narrowed its consideration to 17 comparisons considered reliable (Emans et al. 1991). A statistical analysis that appears valid, compared NOECs from multi-species tests with those from single-species experiments using similar species. The relationship looks reasonable (Fig. 5). Their conclusion: "[T]here seems to be no reason to believe that organisms differ in sensitivity under field and laboratory conditions."

A second objective of Emans et al. "was to study whether ecosystems can be protected by setting a `safe' value that is derived from [single-species] NOECs by extrapolation." Their conclusion: "With reservations, due to this paucity of data, it is concluded that single-species toxicity data can be used to derive "safe" values for the aquatic ecosystem." An earlier version of this review came from Okkerman et al. (1990).

Figure 5. Comparison of results of multiple-species toxicity tests with those from single-species tests (Emans et al. 1991).
This regression of 17 multi-species experiments on single-

species tests is for similar or related species and corresponding effects.

4.6 Review of the AQUAMIN interactive bibliography

The very first task initiated by the author under this contract was an examination of the computerized data-base of environmental reports on mines (AQUAMIN). Almost all of these dealt with Canadian mines, and most were part of the "grey literature" (i.e. not journal articles or numbered reports in a recognized technical series). Unfortunately the list of 734 reports did not appear to provide any studies that could be directly used as validations, in this chapter. At least four of the reports in AQUAMIN contained information that might have been analyzed to provide validations, but the authors had not attempted to provide a focused comparison; that is understandable because the investigations had not been initiated with such a goal.

To be useful for the present review, a report listed in AQUAMIN would have to contain three things: (a) sublethal toxicity tests on the effluent; (b) studies of populations or communities in the receiving water; and (c) a focused comparison by the author, of the results from (a) and (b). AQUAMIN yielded 27 reports when it was searched for (a) toxicity tests in the lab, combined with (b) field studies of macroinvertebrates[9]. Some key descriptions of those 27 reports are given in the Appendix of this chapter. The AQUAMIN indexing indicated that only four of those 27 reports had sublethal or chronic [unlikely] toxicity tests, necessary for the present review of validation. The other 23 had lethal tests and were therefore not useful for the present review. The four reports with sublethal data were numbers 226, 271, 289, and 495 of the AQUAMIN list. They appeared to be good technical studies and had been well-rated in the AQUAMIN tabulation.

Copies of two of the four reports mentioned above were provided by members of the AETE committee. The two reports contained work that was technically well done, but one of them (# 226) had only lethal tests, not sublethal ones as indicated by AQUAMIN. The other (# 289) had information suitable for lab/field comparisons, but cohesive validations had not been done. The report had some general statements on results of sublethal toxicity tests, compared to effects on the receiving community which seemed somewhat milder than might be expected from the toxicity. However, there was no overall or comprehensive comparison of effect-concentrations in the lab with those in the field.

It might be possible to analyze the data in reports such as AQUAMIN # 289, to yield validations. However any re-examinations or re-analysis of data from reports would be time-consuming and far beyond the resources of the present contract. If successful, the results of such re-analysis might provide *comparisons* of toxicity results with effects in the receiving community. There would be no hope, however, of obtaining rigorous *scientific validations*, for the simple reason that validation was not a goal of the work nor a factor in designing it.

About ten other reports on environmental studies at mines were provided by members of the

---

[9] A search of AQUAMIN for "toxicity tests in the laboratory" should obtain all reports which studied the toxicity of effluent. A search for "populations/communities of macroinvertebrates" should obtain all broad field evaluations. Although there could be a good community study which focused on algae, say, or fish, limiting the search to macroinvertebrates did not appear to miss important reports. Examining the list of all 83 reports which included laboratory toxicity, indicated that useful reports had been captured by the first (smaller) search.

AETE committee. All of them were interesting studies in their own right, some of them very perceptive investigations. Most of these reports, however, did not prove useful as examples of field validation of sublethal tests on effluent. None of them had a formal validation or comparison of that nature. All but one of the reports lacked one or other of the two components necessary to make such a comparison. The report that had both components (Times Ltd. 1994) has been described in section 4.1.5 and used in the tabulation of the present review.

4.7 Literature references that were not oriented to the present review

A few publications from the scientific literature were selected as relevant by title but proved to be more oriented to other topics. They are briefly described here, in part because there are some conclusions of general interest, and in part because mention of the contents might save some time for any readers who were searching the literature in the future.

An *in situ* dosing of mixed algal colonies in a stream could be considered a cosm-based "field" study (Lewis et al. 1993). However, the lab single-species values were taken from the literature and proved too highly variable for usefulness. A paper by Clark et al. (1987) dealt with very rapid toxic events, in particular one-time insecticide spraying of an estuary.

A couple of studies turned out to be concerned with sediment contamination. Burmaster et al. (1991) obtained contradictory findings, and recommended: "In understanding the ecological effects of toxic chemicals, there is no substitute for field work." An attempt to develop a mesocosm for marine sediments (Moverley et al. 1995) met variable results: "further research needs to be done ... before mesocosms can be fully developed as routine monitoring tools."

4.8 Conceptual publications

Several publications gave guidance on validation studies in the field or in multi-species communities. Brief mention here is supplemented by notes in the annotated reference list.

Advice on designing validation studies with natural communities is mentioned here, but detailed review is in chapter 3. Sanders (1985) provided excellent general advice on designing validation programs. Cairns (1986, 1988a) listed three desirable phases in the process of scientific validation of laboratory toxicity tests. OECD (1992) presented good principles as recommendations from a workshop. Livingston and Meeter (1985) asked, what were the criteria for verification of laboratory and field results, and answered in complicated fashion.

Cosms and multiple-species tests were topics of advice from several publications. Crane (1985) offered the opinion that model ecosystem tests have not been fruitful. He said that evaluating pesticides in mesocosm ponds cost several million dollars, but the results had no greater sensitivity, predictive power, or interpretability, than cheaper lab tests with single species. *Au contraire,* said Cairns and McCormick (1991), who defended and justified microcosm tests using microbial communities. Lewis (1990) urged more multi-species algal tests, since there were as yet few field validations.

A book on multispecies toxicity testing (Cairns 1985) provided a variety of chapters, many of them academic. Persoone and Janssen (1992) concluded from a wide-ranging review that NOECs from single-species tests related "relatively well" to the highest no-effect levels for field populations, and that prediction improved as more species were included in the lab tests.

4.9 Tabulation of validation studies and conclusions

Table 1 gives a condensed view of the useful parts of the information reviewed in section 4. The studies in the literature are varied in their nature, no one is exactly the same as another, and many of the conclusions are qualitative, or qualified rather than clear-cut. Accordingly I will not attempt any formal arithmetic analysis of the degree of agreement. An approximate tally is given in the conclusions (next section).

Conclusions

(1) Most of the validation studies were not carried out in a rigorous scientific manner designed to test an hypothesis of "predictability". They were mostly general correlations of (a) a single-species toxicity test on an effluent or on a sample of the receiving water, with (b) changes in the community in the affected receiving water, or in an artificial stream, pond, or other mesocosm. Some comparisons were formal statistical correlations, others were *ad hoc* mathematical comparisons, and many were only subjective or graphical comparisons. One study, however, did a purely objective, predictive comparison of 43 cases (Eagleson et al. 1990).

(2) Despite the lack of rigorous scientific method in most validation studies, most technical people who read the publications would probably be convinced that conclusions drawn by the authors were reasonable judgments on agreement or disagreement.

(3a) Among the reports reviewed, there was a strong balance of agreement between the results of single-species laboratory tests, and observed effects on communities (Table 1).

(3b) A rough tabulation might be made. Of the 29 studies listed in Table 1, 14 give a "yes" answer for reasonable agreement, and 5 give a "no" because lab tests did not agree with the field. Two studies must be considered marginal and not useful in this comparison, because there were no effects in the field study, and such lack of effect could have various meanings. Eight other studies are not useful in the comparison because they involved lethality not sublethal effects, because they were literature searches not experimental projects, or simply because the project was not completed successfully. Thus there are 14 studies in agreement, and 5 in disagreement, or 74% of the usable studies showing reasonable correlation of lab and field.

One of the studies counted above as showing agreement, involved three comparisons with different chemicals (the EEC study). Another involved comparisons with 43 effluents and field surveys of which 88% or 38 comparisons were in agreement (the predictive study in North Carolina by Eagleson et al.). If these breakdowns were used, the total tally would be 53 cases in agreement and 10 in disagreement, for a success rate of 84%.

Table 1. Listing and diagnosis of validation studies reviewed in sections 4.1 to 4.3 and 4.5. Order of studies is same as in the text; first-column identification assists locating in the reference list. Fh = sublethal test with fathead minnows; C = with *Ceriodaphnia*; var. = with various tests. (Parentheses indicate an item that is not directly useful for this review.)

Brief identification	No. of items	Single-species lab tests			"Field" tests		Agreement?
		Sub-lethal	Eff-luent	Amb-ient	River, strm.	Art. cosm	
U.S. EPA, Geckler	1	Fh		yes	yes		Yes, both affected
U.S. EPA, eight validation studies, by Mount, Norberg-King, et al.							
(Ohio, chemical resin)	1	Fh C	yes	yes	yes		(Yes but both non-toxic)
(W. Virginia)	1	Fh C	no	yes	yes		(Partly. plankton but not invert.)
Ottawa River	1	Fh C	yes	yes	yes		Yes
Alabama	1	Fh	yes	yes	yes		Yes, but no single test overall
Naugatuck River	1	Fh C	yes	yes	yes		Good
(Baltimore)	1	Fh C		yes	inadequate		(Not known)
Ohio River	1	Fh C	no	yes	yes		Ceriodaphnia reasonably good
Oklahoma	1	Fh C	yes	yes	yes		Yes, predicted worst station
Kentucky, Birge	1	Fh	yes	yes	yes		Yes
N. Carolina, Eagleson	43	C	yes		yes		88% agreement, true prediction
N. Dakota mining, Times	1	Fh C	yes		yes		Disagreed, no effect in stream
Virginia, Pontasch	1	C	yes		yes		Good, (1.7% versus 2.0%)
Colorado, Davies & Woodley	1	var.	no	no	yes		No, fish harmed at "safe" level
(Colorado, Clements ...)	1	C, unsatisfact.			unsatisfactory		(Not known)
(Anti-pollutant, Sprague)	1	lethality			yes		(Agreed, but not sublethal)
(Pulp mills, Robinson)	11	Fh C	no	yes	literature only		(General agreement)
EPA channels, PCphenol	1	var.	chemical			yes	Close, slight effect on fish
EPA channels, selenium	1	"safe"	dosing			yes	No, extra oral intake of toxicant
(EPA channels, *p*-cresol)	1	lethal	of the			yes	(Good, but not sublethal)
Ethers, Crossland	1	Daph	controlled			yes	Yes, reasonable
(LAS, Fairchild)	1	Fh +	cosms.			yes	(Good, but field effect absent)
Diflubenzuron, Hansen	1	var.	"			yes	Adequate
EEC, 5 labs	3	var.	"			yes	Agreement stated
EEC, 5 labs	1	var.	"			yes	No, 10-fold error
Atrazine, Larsen	1	algae	"			yes	"Excellent"
Copper, Moore & Winner	1	C	"			yes	No, some field effect at "safe"
(Literature review, Slooff)	many	var.	no	no	various		(Good general relationship)
(Literature review, Emans)	17	var.	no	no	various		(Can predict, lab to ecosystem

Annotated references

Annotations follow many references, particularly the ones used in section 4, concerning validation. The annotations give additional details, direct quotations from the authors, or guidance to readers on whether the original might be worth retrieving.

Alabaster, J.S. and R. Lloyd 1980. Water quality criteria for freshwater fish. Butterworth Pub. Inc., London, U.K. and Woburn, Mass. U.S.A.

Arthur, J.W., J.A. Zischke, K.N. Allen and R.O. Hermanutz 1983. Effects of diazinon on macroinvertebrates and insect emergence in outdoor experimental channels. Aquatic Toxicol. 4: 283-301.
This study did not carry out laboratory tests for direct comparison with the studies in the channels. It is an example of using artificial streams. Relative tolerance of invertebrates were estimated and listed.

Birge, W.J., J.A. Black, T.M. Short and A.G. Westerman 1989. A comparative ecological and toxicological investigation of a secondary wastewater treatment plant effluent and its receiving stream. Environ. Toxicol. Chem. 8: 437-450.
This is an excellent comparison of pollution impacts in a stream with results of fathead minnow embryo-larval tests. It is clear that the research was carried out with a high standard of quality. The comparison used 8-day embryo-larval toxicity tests with mortality as the main criterion of effect. These tests were run on dilutions of effluent and also on samples of receiving water ("ambient" tests). Agreement was extremely good between toxicity tests and surveys of invertebrates and fish (as well as chemical conditions).

"Toxicity measurement reported for the receiving water system compared closely with independent ecological parameters, particularly species richness and diversity (H) of macroinvertebrates. The correlation coefficients for these data against percent embryo-larval survival were 0.96 and 0.93, respectively ... Thus it was concluded that ... results of the ... 8-d fathead minnow embryo-larval tests performed in the receiving waters provided reliable estimates of ecological impact."

Some numbers might be cited. At station 4, ambient tests showed 14% mortality of fathead minnow larvae, significantly higher than control values of 7 and 10% mortality. At station 4, the concentration of effluent in the stream was 53%, and dissolved oxygen was only 5.9 mg/L compared to control values of 9.1 and 9.7 mg/L. There were 10 species of fish, similar to control values of 8 and 11. However, the invertebrates showed only 22 species compared to control values of 30 and 34, and diversity index of only 2.5 compared to 3.3 and 3.7 (a value of 3.0 and above suggests clean water). Station 5 had incomplete data. Station 6 was the first one to show no difference from control values in either the toxicity tests or field data. Mortality of larvae was 8%, concentration of effluent was 33%, oxygen was 9.3 mg/L, fish species numbered 11, and there were 29 species of invertebrates with a diversity index of 3.8.

The effluent tests also showed close agreement with field observations. The tests estimated a lethality threshold (1% mortality) at 36% effluent concentration, almost the same as the 33% concentration present at station 6.

Burmaster, D.E., C.A. Menzie and J.S. Freshman 1991. Assessment of methods for estimating aquatic hazards at superfund-type sites: a cautionary tale. Environ. Toxicol. Chem. 10: 827-842.
A complex program of studies, mostly involving sediment contamination. Toxicity tests were of lethality from sediment, and do not have direct bearing on the present review. Some general advice arises.

"[T]he theoretical calculations,laboratory experiments and field studies ... yielded a patchwork of confirmatory and contradictory findings. ... In understanding the ecological effects of toxic chemicals, there is no substitute for field work."

Cairns, J. Jr. (ed.) 1985. Multispecies toxicity testing. Soc. Environ. Toxicol. Chem. Special Pub. Series, Pergamon Press, Elmsford, N.Y..
Some chapters from this book are cited. The book itself is a comprehensive study of this topic for its time, although some parts are rather academic or theoretical. The conference which generated this book led to a set of conclusions, given at the back of the book. Among them is: "Multispecies tests will remain most useful as research tools, not as tools for screening large numbers of chemicals for their relative toxicity."

Cairns, J. Jr. 1986. What is meant by validation of predictions based on laboratory toxicity tests? Hydrobiologia 137: 271-278.
Cairns, J. Jr. 1988a. What constitutes field validation of predictions based on laboratory evidence? P. 361-368 *in*: W.J. Adams, G.A. Chapman and W.G. Landis (eds). Aquatic Toxicology and Hazard Assessment: Tenth Volume. Amer. Soc. Testing and Materials, Philadelphia, Pa., ASTM STP 971.
Similar points in the two publications, first indirectly quoted from Persoone and Janssen (1992).

There should be three phases in the scientific validation process for laboratory toxicity tests.
Stage 1 should show that a species present in the receiving water has an threshold of effect that is reasonably close to the threshold for the test species, or else is more tolerant than the test species.
Stage 2 should show that an array of species from a variety of taxonomic groups and trophic levels do not suffer deleterious effects at the predicted no-effect levels derived from the laboratory work.
Stage 3 should show that important attributes of the community or ecosystem such as energy flow or nutrient cycling are not impaired at the predicted no-effect levels.

Cairns, J. Jr. 1988. Politics, economics, science -- going beyond disciplinary boundaries to protect aquatic eco systems. P. 1-16 *in* M.S. Evans (ed.). Toxic contaminants and ecosystem health: a Great Lakes focus. Ad vances in Environ. Sci. Technol., vol. 21. Wiley, New York, N.Y.

Cairns, J. Jr. and P.V. McCormick 1991. The use of community- and ecosystem-level end points in environmental hazard assessment: a scientific and regulatory evaluation. Environmental Auditor 2: 239-248.
Largely a defence and justification of microcosm tests using microbial communities (bacteria, algae, protozoans, and some other small animals such as rotifers). Some examples and some useful references to validation.

Cairns, J. Jr., E.P. Smith and D. Orvos 1988. The problem of validating simulation of hazardous exposure in natural systems. Proc. 1988 Summer Computer Simulation Conf., p. 448 -454. C.C.

Barnett and W.M. Holmes (eds). Soc. Computer Simulation Internat., San Diego, Calif.
Wide-ranging, but particularly relevant is the attack on the U.S. EPA attempts to "validate" single-species tests (the 8 studies Mount, Norberg et al.). They say that there are two types of validation, (a) a preliminary validation which assesses the similarity between the model (single-species results) and the real system (polluted ecosystem), and (b) predictive validation which deals with the relationship between the model and the system, i.e., how well the model works in practice. EPA was said to have used the first kind of validation, by simply correlating effects in the field with effects in single-species laboratory tests, and that was not *predictive* validation.

One part of the reports of the U.S. EPA involved comparison of toxicity results, say for *Ceriodaphnia*, from (a) a purely laboratory toxicity test of effluent, with (b) a toxicity test of the affected surface water, the so-called "ambient" toxicity tests. As might be expected, there was reasonable agreement in results allowing for dilution, but that comparison was not validation with respect to a multi-species community.

"A second approach to validation ... was to compare the ambient water toxicity data on *Ceriodaphnia* with the field survey data on other related species and measures related to community structure. The study found a correlation between the toxicity tests and the structure measures. What this means is that the pattern in the toxicity test is similar to the pattern in the field. It does not mean, contrary to the statement that appears in the Executive Summary, that `effluent and ambient toxicity tests are accurate predictors of receiving water impact'. ... *A good correlation between the toxicity data and the field data might only mean that the two data sets have a similar pattern, not that one might accurately predict the other!* While this might appear to be only semantics, it is important since the management at EPA is led to believe that good predictions about ecological impact using this methodology may be obtained."

The authors point out that to validate the use of the single-species tests as accurate predictors, the experiment would have to (1) carry out the single-species toxicity test, (2) predict the field data, *then* (3) collect the field data, and (4) assess the accuracy of the prediction. [The author of the present report would add: (2a) set up criteria for agreement versus disagreement.

Clark, J.R., P.W. Borthwick, L.R. Goodman, J.M. Patrick Jr., E.M. Lores and J.C. Moore 1987. Comparison of laboratory toxicity test results with responses of estuarine animals exposed to Fenthion in the field. Environ. Toxicol. Chem. 6: 151-160.
Of limited use for this review since it deals only with acute toxicity. Laboratory tests of acute lethality were done for marine crustaceans and a fish. Results agreed with findings when caged animals were held in estuarine waters sprayed with the insecticide. For short events of <24 hours, lab data for pulse exposures were necessary for good correlation.

Clements, W.H., J.L. Farris, D.S. Cherry and J. Cairns Jr. 1989. The influence of water quality on macroinvertebrate community responses to copper in outdoor experimental streams. Aquatic Toxicol. 14: 249-262.
Not of direct interest for this review. They exposed communities in artificial streams in two kinds of stream water, dosed with copper. General agreement with results from natural communities affected by metal pollution. This showed that the artificial streams behaved in similar fashion to real streams.

Clements, W.H. and P.M. Kiffney 1994. Integrated laboratory and field approach for assessing impacts of heavy metals at that Arkansas River, Colorado. Environ. Toxicol. Chem. 13: 397-404.
Of some use in the present review, since they compared ambient sublethal toxicity tests of *Ceriodaphnia* reproduction with instream communities of macroinvertebrates (and other less relevant things). There

was not good agreement. Macroinvertebrate populations differed somewhat in autumn and spring, but these field observations indicated moderate effects on both of two downstream stations. *Ceriodaphnia* reproduction was less affected in water of the first downstream station (23 neonates produced compared to control values of 26 to 34) than in the second downstream station (6 neonates) where effects should have been less. Tests in another season indicated severe effects on *Ceriodaphnia* at the first downstream station (1 neonate). The control of the *Ceriodaphnia* test was almost as badly affected (11 neonates) as some tests using water from polluted locations. I conclude a general disarray of findings, rather than any disagreement or agreement between field and lab. The authors end by recommending "an integrated approach" i.e. several methods of investigation.

Cooper, W.E. and R.J. Stout 1985. The Monticello experiment: a case study. P. 96-116 ***in*****: J. Cairns Jr. (ed.). Multispecies toxicity testing. Pergamon Press, New York, N.Y.**
They tested *p*-cresol in laboratory and in full-size, outdoor artificial streams, in a very large, multi-specialist, expensive experiment. Single-species acute toxicity tests in the laboratory tested 3 species of fish, the crustaceans *Daphnia magna* and *Hyalella azteca*, and a damselfly. These indicated lethal thresholds in the vicinity of 10 to 20 mg/L for the fish, and 2 to 5 mg/L for the crustaceans. The test streams were dosed with 8 mg/L for 1, 2, or 4 days. The species, numbers, and biomass of species in the streams were measured during and after exposures, and "community" variables were species diversity and evenness, community metabolism, and degradation of added packages of leaves ("leaf-packs"). The detailed results cannot be summarized here, but there appeared to be reasonable agreement between the single-species lab tests and the stream ecosystem results.

The authors concluded such agreement. "Hypothesis I: The transfer of laboratory acute toxicity tests to a field situation is possible without serious distortion. Conclusion I: The acute toxicity tests with fathead minnows, largemouth bass, smallmouth bass, damselflies, and amphipods produced estimated survivorship rates of exposure to *p*-cresol that were consistent with the results of the field experiments."

Hypothesis and conclusion number two said that the "community" variables "indicated the same type of ecological impacts on macroinvertebrates as did the single species analyses." The major impacts in the streams were actually indirect, through effects on photosynthesis and respiration of aquatic plants, but still the general correspondence of concentration/effect prevailed.

Crane, M. 1995. Is there a place for ecology in ecotoxicology. ***SETAC News*** **March 1995, 19-20.**
In this letter of opinion to the editor, Crane says that higher theories of ecology have not proved useful in aquatic toxicology, and that model ecosystem tests have not been fruitful. "... the current consensus among ecotoxicologists who have performed such studies is that model ecosystems are not particularly useful tools for looking at ecological interactions."

Concerning mesocosm tests in ponds to evaluate pesticides according to a method of U.S. EPA, he states: "Such studies cost several million dollars to perform, but the results obtained from them have shown no greater sensitivity or predictive power, and certainly no greater interpretability, than considerably cheaper laboratory tests with single species."

CCREM 1987. Canadian water quality guidelines. Canadian Council of Resource and Environment Ministers, Task Force on Water Quality Guidelines. Environment Canada, Ottawa, Ontario.
A complete and internationally recognized review and recommendation of water quality objectives for many chemicals.

Crossland, N.O. and G.C. Mitchell 1992. Use of outdoor artificial streams to determine threshold toxicity concentrations for a petrochemical effluent. Environ. Toxicol. Chem. 11: 49-59.
A useful study. The authors did their own sublethal lab tests and then used metal artificial streams to assess invertebrate populations, species, drift and feeding rates. This was a well-run test with measured concentrations at constant level. The toxic effluent was a mixture of chlorinated ethers from the organic fraction of a petrochemical manufacturing plant. The lab single-species NOEC was 1.0 mg/L for growth of *Daphnia magna*, while the multispecies NOEC was 0.44 mg/L for feeding of *Gammarus pulex*, reasonable agreement since the two values differed by only 2.4-fold.

This entire issue of the journal was devoted to 10 papers and an editorial on mesocosms. This is the only paper of direct interest here for its lab/field comparison.

Davies, P.H. and J.D. Woodling 1980. Importance of laboratory -derived metal toxicity results in predicting in-stream response of resident salmonids. *In:* J.G. Eaton, P.R. Parrish and A.C. Hendricks (eds.). Aquatic Toxicology. American Society for Testing and Materials, Philadelphia Pa. ASTM STP 707, p. 281 -299.
This was a study of the effects of metals from mine discharges in a Colorado river. It compared historic values from toxicity tests in the laboratory with lethality tests with fish *in situ*, and field surveys of fish. Thus it is not exactly comparable to the present interest of evaluating sublethal tests, but provides some information on agreement of established laboratory criteria for toxicity, with field observations.

"Toxicity results from *in situ* bioassays, ... agree closely with laboratory-derived toxicity findings ..." When concentrations of zinc, copper and cadmium rose above acutely toxic values as determined in historic laboratory tests, "resident brown and brook trout were severely impacted " and "acute toxicity occurred during the August *in situ* bioassay ..." "Stations farther downstream, where the concentrations of metals decreased to [estimated safe levels as determined in historic laboratory tests] for brook trout, showed a marked reduction in the number of resident salmonids ..."

The last statement may be taken as signifying that the field survey showed greater sensitivity of fish than did single-species sublethal tests in the laboratory, i.e. disagreement of the two approaches.

Dickson, K.L., W.T. Waller, J.H. Kennedy and L.P. Ammann 1992. Assessing the relationship between ambient toxicity and instream biological response. Environ. Toxicol. Chem. 11: 1307-1322.
These authors were contracted by U.S. EPA to carry out a meaningful statistical analysis of the eight EPA attempts at validating toxicity tests as predictors of instream conditions. The authors carried out a canonical analysis of the data, starting afresh with the observations from the eight reports.

The results of their analysis are somewhat reduced for purposes of the present review because they ended up comparing the instream community surveys with "ambient" toxicity tests (i.e. instream tests, in which water from the stream was tested for its toxicity). The focus of the present review is on effluent tests, and a comparison with such tests would have been more directly relevant. However, there is little question that any toxicity measured in ambient tests of EPA would have been derived from effluents, and in fact it would be straightforward to move mathematically from effluent toxicity to ambient toxicity, by factoring in dilution. Therefore the analyses of Dickson et al. are relevant to the present review.

The canonical analysis is complex and cannot be summarized here in a few words. The authors characterize the canonical approach as analogous to linear regression applied to groups of response variables handled simultaneously. We can accept the output of the analyses, which is helpful. The following table represents one output. It is a contingency table resulting from an application of 94-95 percentile cutoffs to the data in a binomial model; this approach would minimize the chance of erroneously predicting that a site was not affected when it actually was affected (i.e. little chance of calling a polluted site clean).

	Instream survey finds impact	**Instream survey finds no impact**
Lab test predicts impact	Agreement. 69 cases (= 86%)	Disagreement. 6 cases (= 8%). Lab test is apparently over-protective
Lab test predicts no impact	Disagreement. 2 cases (= 2%). Lab test is apparently under-protective	Agreement. 3 cases (= 4%).

The point here is that there was 90% agreement of classification. Cutoffs of 5-5 percentiles could have been used to minimize the possibility of error in predicting that a "clean site was polluted". The numbers from that analysis are not essential here, but again the point is that the agreement was 85%.

The authors conclude from those results "that the relationship between ambient toxicity and instream biological response is strong, regardless of which misclassification error is of greatest concern." [Mistakenly classifying a polluted site as clean would presumably be of greatest concern to a government agency, while classifying a clean site as polluted would be of concern to an industry that was discharging effluent to that site.]

The authors make other useful statements. They say that their analysis "demonstrated that statistically significant relationships between ambient toxicity and instream impact existed for the examined data sets". The variables in the toxicity tests found to be most useful were neonate production of *Ceriodaphnia*, and dry weight of fathead minnow larvae. Useful instream variables were numbers of species of fish and of invertebrates. They point out that there were many confounding factors in the EPA studies (e.g. unknown spills) that weakened the analysis, but still a relationship emerged.

This same paper also analyzed two other detailed studies which are not dwelt upon in the above summary, but I have included the findings in summaries of the original studies, by Birge et al. (1989) and Dickson et al. (1989), where they appear in this annotated reference list.

Eagleson, K.W., D.L. Lenat, L.W. Ausley and F.B. Winborne 1990. Comparison of measured instream biological responses with responses predicted using the *Ceriodaphnia dubia* chronic toxicity test. Environ. Toxicol. Chem. 9: 1019-1028.
This is a most useful paper. There is a lot of information, the procedures are clear-cut and objective, and the results are convincing. The authors present the result of 43 studies of complex effluents in North Carolina and the surface waters into which they were discharged. There were 9 industrial effluents, the rest were municipal but usually included industrial components. Toxicity was assessed by the 7-day *Ceriodaphnia* method, and it was then *predicted* whether there should be effects in the waterbody (i.e. the prediction was made in the appropriate sequence). The decision was based on whether the predicted "safe" concentration was higher or lower than a river concentration defined as the permitted discharge

from the plant divided by the 7Q10 (the lowest 7-day average flow expected every 10 years). If the laboratory "safe" level was at a higher concentration, then no effect was predicted in the river. A survey of macroinvertebrates was made in each waterbody and an effect was decided by normal criteria of species and numbers. The overall agreement of prediction and field finding was 88%, distributed as in the following table.

	Instream survey finds impact	**Instream survey finds no impact**
Lab test predicts impact	Agreement. 29 cases (= 67%)	Disagreement. 3 cases (= 7%). Lab test is apparently over-protective
Lab test predicts no impact	Disagreement. 2 cases (= 5%). Lab test is apparently under-protective	Agreement. 9 cases (= 21%).

The authors are perhaps conservative in their conclusion: "These data suggest that the use of effluent toxicity testing results as a regulatory tool is effective and appropriate."

EEC (European Economic Community) 1992. Development and validation of methods for evaluating chronic toxicity to freshwater ecosystems. Final report (summary). EEC, Environ. Res. Programme, Assessment of Risk Associated with Chemicals (Ecotoxicology).
As stated in the title this is a final report, of 5 laboratories which carried out research on what was intended to be a unified project. They developed some sublethal tests and did validation experiments in streams and ponds. Unfortunately the report is very fragmented with detailed results presented separately by each laboratory. For example one lab reported the work on pond mesocosms, another the work on stream mesocosms, while the others reported the tests they had developed and the laboratory toxicity tests. Dozens or hundreds of endpoints are reported.

There is a one-page summary, but it has some very general statements (some quoted below) which are not very satisfactory because they are not supported by listing the numerical data used for the conclusions. Many publications have come out of this large project already and more were reported as underway. It is to be hoped that there will be a comprehensive and documented summary of what it all means. I will report here, some of what they did, and the general conclusions made on validation. I am unable to carry out a complete analysis of the data reported in any reasonable time-frame.

The single-species laboratory tests included:
- 10-day growth tests with two planktonic algae;
- 5-day growth tests with two protozoans;
- 10-day life-cycle test with a rotifer;
- 10 and 28-day growth tests with macroinvertebrates.

The chemicals tested were copper, lindane, atrazine, and 3,4-dichloroaniline.

"Field experiments to determine threshold concentrations for effects of the 4 reference chemicals were carried out in artificial streams (28-36 day tests) and in enclosures in ponds (39-62 day tests). The data from these laboratory chronic toxicity tests provided threshold concentrations which were protective of the poind and stream ecosystems except in the case of 3,4-dichloroaniline where effects were observed in the field at concentrations approximately 1/10 of the lowest laboratory threshold value."

Emans, H.J.B., E.J. v.d. Plassche, J.H. Canton, P.C. Okkerman and P.M. Sparenburg 1991. Validation of some extrapolation methods used for effect assessment. Environ. Toxicol. Chem. 12: 2139-2154.

This is a massive literature study, most useful for the present purposes for its comparison of "safe" levels derived from single-species tests with those from multi-species tests. In general, they compare the two categories by using similar species and effects. After eliminating unreliable studies or those with no paired comparison, they were left with 17 comparisons of single- versus multi-species. It is a complicated paper to follow. Much of it is devoted to evaluating several techniques of extrapolation from effect-levels to no-effect levels, something which is not of immediate interest for the present review.

Their *first objective* "was to study whether there are relevant differences in NOECs derived from [multi-species] and [single-species] experiments for similar or related species and corresponding effect parameters." Their conclusion was: "[T]here seems to be no reason to believe that organisms differ in sensitivity under field and laboratory conditions. When species tested in [multi-species] experiments were compared with similar or related species in [single-species] experiments, for corresponding effect parameters and exposed to equal concentrations, they appeared to be equally sensitive. This result was supported by statistical analysis of the available data by means of model II regression and a *t* test for paired comparisons."

Their *second objective* "was to study whether ecosystems can be protected by setting a "safe" value that is derived from [single-species] NOECs by extrapolation." Their conclusion: "With reservations, due to this paucity of data, it is concluded that single-species toxicity data can be used to derive "safe" values for the aquatic ecosystem."

This publication can also be used as a source of multi-species toxicity experiments considered "reliable", although most of the projects identified do not have any associated laboratory tests with single species for the purpose of predicting the multi-species result.

Environment Canada 1992a. [Written by D.J. McLeay and J.B. Sprague.] Biological test method: test of reproduction and survival using the cladoceran *Ceriodaphnia dubia.* Environment Canada, Conservation and Prot., Ottawa, Ontario. Rept EPS 1/RM/21.

Environment Canada 1992b. [Written by J.B. Sprague and D.J. McLeay.] Biological test method: test of larval growth and survival using fathead minnows. Environment Canada, Conservation and Prot., Ottawa, Ontario. Rept EPS 1/RM/22.

Environment Canada 1992c. [Written by D. St-Laurent, G.L. Stephenson, and K.E. Day] Biological test method: microplate growth inhibition test with alga *(Selenastrum capricornutum)*. Environment Canada, Conservation and Prot., Ottawa, Ontario. Rept EPS 1/RM/25.

Environment Canada 1992d. [Written by M.R. Gordon, D.J. McLeay and J.B. Sprague.] Biological test method: toxicity tests using early life stages of salmonid fish (rainbow trout, coho salmon, or Atlantic salmon). Environment Canada, Conservation and Prot., Ottawa, Ontario. Report EPS 1/RM/28

Fairchild, J.F., F.J. Dwyer, T.W. La Point, S.A. Burch and C.G. Ingersoll 1993. Evaluation of a laboratory-generated NOEC for linear alkylbenzene sulfonate in outdoor experimental streams. Environ. Toxicol. Chem. 12: 1763-1775.
For a series of laboratory tests with 10-day-old fathead minnows, including a 28-day study of mortality and growth, the lowest TOEC was 0.42 mg/L of the detergent. Seven-day exposures of the amphipod *Hyalella azteca* showed a lowest TOEC of 1 mg/L; that could be considered a "sub-acute" effect on mortality. A 45-day run of 50-m long experimental streams at 0.36 mg/L of LAS showed no effect on periphyton, macrobenthos, or biological processing of dead leaves. This indicates, at least, no disagreement between the single-species tests and the mesocosm experiment, since the threshold of effect in the laboratory (0.42 mg/L) was protective of biota in the stream exposures at approximately the same concentration (0.36 mg/L). However, the absence of effect does not provide a means of formally correlating lab and field.

Geckler, J.R., W.B. Horning, T.M. Neiheisel, Q.H, Pickering, E.L. Robinson and C.E. Stephan 1976. Validity of laboratory tests for predicting copper toxicity in streams. U.S. Environmental Protection Agency, Ecol. Res. Ser., EPA -600/3-76-116. Washington, D.C.
A major research project. They dosed a real creek (Schayler Run) with copper for two years and studied fish and invertebrates and other aspects. The streamwater contained sorbing substances from an upstream sewage treatment plant, and when the streamwater was used as dilution water in the lethal and sublethal lab tests, results agreed with findings in the stream. (Lab tests with clean water did not agree because there was no sorbing substance present and the copper showed its unhampered toxicity.)

The big finding was that avoidance responses of wild fish were more sensitive than physiological effects. In sections of the stream in which reproducing populations of fish were possible, there were none because they had moved downstream. Some species *could have* spawned in mildly polluted sections of the stream, but did not; their avoidance reaction made them flee downstream until they reached the lower barrier fence, where they spawned. The same species confined in cages in the mildly polluted section, spawned in the cages since they could not flee; their (physiological) ability to reproduce at those locations agreed with the laboratory predictions.

Hansen, S.R. and R.R. Garton 1982. Ability of standard toxicity tests to predict the effects of the insecticide diflubenzuron on laboratory stream communities. Can. J. Fish. Aquat. Sci. 39: 1273-1288.
They did five sublethal single-species tests in the laboratory, evaluating mortality, growth and fecundity. They compared with biomass and diversity of communities in artificial streams in the laboratory, with communities established for 3 months and then dosed for 5 months.

"The single-species tests adequately predicted the concentrations of diflubenzuron which affected these stream communities; the most-sensitive test species, insects and crustaceans, were up to an order of magnitude more sensitive than the observed community effects." The single-species tests did not predict the particular mechanisms of effect in the streams.

Hedtke, S.F. 1984. Structure and function of copper-stressed aquatic microcosms. Aquatic Toxicol. 5: 227-244.

Herbert, D.W.M. 1965. Pollution and fisheries. Ecology and the Industrial Society, 5th Symposium British Ecol. Soc., 173-195. Blackwell Scientific, London.
A somewhat indirect study involving a detour through chemical concentrations. The toxic units method

described under Lloyd and Jordan (1964) was used to predict to an actual river. Caged fish were renewed in the river and daily assessments of mortality were made. Predictions were made on the basis of toxic units of the mixture of chemicals measured in the river. Herbert notes that 83% of the predictions of lethality to fish in the river were correct (Figure 3).

Hermanutz, R.O., K.N. Allen, T.H. Roush and S.F. Hedtke 1992. Effects of elevated selenium concentrations on bluegills *(Lepomis macrochirus)* in outdoor experimental streams. Environ. Toxicol. Chem. 11: 217-224.
This paper is an example showing that it would not be wise to depend *only* on single-species laboratory tests, for some toxic substances. Previous laboratory toxicity tests had estimated that 26 μg/L of selenium was a "safe" level for aquatic life, and that concentration had been issued as a water quality criterion by the U.S. EPA. Findings in the field had indicated effects at lower concentrations. This one-year exposure of fish in artificial streams confirmed the field observations, that even 10 μg/L had appreciable sublethal effects on fish (growth, survival of young, edema and other internal ills). The greater sensitivity came from intake of selenium from food as well as water, while only the water route had been included in the single-species laboratory tests.

Kaesler, R.L., J. Cairns Jr. and J.S. Crossman 1974. Redundancy in data from stream surveys. Water Res. 8: 637-642.

Klapow, L.A., and R.H. Lewis 1979. Analysis of toxicity data for California marine water quality standards. J. Water Pollut. Control Fed. 51: 2054 -2070.

Kuiper, J. 1981. Fate and effects of cadmium in marine plankton communities in experimental enclosures. Marine Ecol. Progr. Ser. 6: 161-174.

Larsen, D.P., F. deNoyelles, F. Stay and T. Shiroyama 1986. Comparisons of single species, microcosm and experimental pond responses to atrazine exposure. Environ. Toxicol. Chem. 5: 179-190.
They compared single-species tests, microcosm, and experimental pond, to examine effects of atrazine on 8 species of algae. Good replications were reported, with 50% inhibition of photosynthesis, respiration, and algal biomass occurring in all three systems at atrazine concentrations within the range 100 to 155 μg/L.

Lewis, M.A. 1990. Are laboratory-derived toxicity data for freshwater algae worth the effort? Environ. Toxicol. Chem. 9: 1279-1284.
No particular data in this paper. Author comments that there are few field validations of algal tests and urges multi-species algal tests.

Lewis, M.A., C.A. Pittinger, D.H. Davidson and C.J. Ritchie 1993. In situ response of natural periphyton to an anionic surfactant and an environmental risk assessment for phytotoxic effects. Environ. Toxicol. Chem. 12: 1803-1812.
This is not a very useful comparison. They did *in situ* dosing of mixed algal colonies in a stream, and found a TOEC of 3.3 mg/L; that could be considered a cosm-derived "field" effect. However, they depended on the literature for laboratory single-species values, and they were extremely variable with

species, from 0.9 to 120 mg/L.

Livingston, R.J. and D.A. Meeter 1985. Correspondence of laboratory and field results: what are the criteria for verification? P. 76-88 *in*: Cairns, J. Jr. (ed.). Multispecies toxicity testing. Pergamon Press, New York, N.Y.
Rather complicated and difficult to understand. Not used in this review

Lloyd, R. and D.H.H. Jordan 1964. Predicted and observed toxicities of several sewage effluents to rainbow trout: a further study. J. Proc. Inst. Sewage Purif. Pt. 2, 183-186.
Classic study of prediction, based on acute lethality. See Figure 4. They determined lethal concentrations of common pollutants, and assumed that in a mixture they would add up in their toxicity. [By the *toxic unit* method. (Measured concentration of toxicant A in river) divided by (Lethal threshold concentration of toxicant A as determined in the lab) = toxic units of toxicant A. The same is done for toxicant B, C, etc., and the toxic units are added up. If they exceed 1.0 the mixture is predicted to be lethal. The same system can be used for sublethal toxic units by having the denominator as the "safe" level or water quality standard.] They studied various effluents, measured chemical pollutants, added up in toxic units to predict whether or not lethal, and the number of toxic units. Samples of effluent were tested for toxicity, the LC50 and actual number of toxic units calculated. Agreement of predicted and observed toxic units was found in 19 cases, for 24 effluents tested. This is an indirect prediction, voyaging to its destination via chemical tests, but provides general supporting evidence of the dependability of predictions from toxicity tests. Some work in rivers is reported by Herbert (1965).

Marcus, M.D. and L.L. McDonald 1992. Evaluating the statistical bases for relating receiving water impacts to effluent and ambient toxicities. Environ. Toxicol. Chem. 11: 1389-1402.
This is an important paper. It judges the U.S. EPA Program on Testing Toxicity of Complex Effluents (CETTP) in eight waterways, which was an attempt to validate sublethal effluent testing as a means of predicting effects in the receiving water. The paper does not have negative comments on the experimental parts of the work (data collection in lab or field), but pretty well demolishes the EPA methods for comparing and analyzing results.

Briefly, Marcus and McDonald conclude that the CETTP studies did not include statistical evaluation of relationships between effluent toxicity and instream effects. Design limitations mean that inferences from the work cannot be applied to other places or other times. There were few statistically significant correlations between pairs of ambient (instream) toxicity tests and variables of the instream communities. The method used by EPA to show correct predictions was "mathematically inappropriate".

The authors seem to be mathematically erudite and apply their own analyses to a limited extent, including canonical correlation analysis. They conclude: "Our results indicate that the short-term chronic test results for ambient waters can potentially provide useful information relating to biological community structure in streams where toxicity is a predominating habitat influence". More specifically, they generalize that "Our analyses indicate that the short-term chronic *Ceriodaphnia* and 7-d fathead minnow tests used to determine ambient toxicity can provide useful information about biological community structure in stream waters where ambient toxicity has a controlling influence".

Behind those statements are more detailed conclusions as in the following examples. For one study (Naugatuck River) a high proportion (>85%) of variability in the ambient (instream) toxicity data was associated with the patterns of variability found in the community taxa data. There were significant correlations between the numbers of taxa in the field sampling, and the ambient toxicity, in 45% of the data sets. More generally, there was >50% relationship between the variables in sets of laboratory toxicity and field data, for 9 of 11 analyses by Marcus and McDonald (and a potentially important relationship in the other 2 sets of data). From this, Marcus and McDonald conclude "that effluent toxicity had important influences in these systems".

McKim, J.M. 1977. Evaluation of tests with early life stages of fish for predicting long -term toxicity. J. Fish. Res. Board Canada 34: 1148 -1154.

McKim, J.M. 1985. Early life stage toxicity tests. P. 58 -85 *in* G.M. Rand and S.R. Petrocelli (eds). Fundamentals of aquatic toxicology. Methods and applications. Hemisphere Pub. Corp., Washington, D.C.

Microbics 1995a. Microtox chronic toxicity test. Microbics Corp., Carlsbad, Calif. 38 p.

Microbics 1995b. Microtox chronic toxicity testing software user's guide. Microbics Corp., Carlsbad, Calif. 69 p.

Moore, M.V. and R.W. Winner 1989. Relative sensitivity of *Ceriodaphnia dubia* laboratory tests and pond communities of zooplankton and benthos to chronic copper stress. Aquatic Toxicol. 15: 311-330.
Of some interest for this review. They did 7-day sublethal *Ceriodaphnia* tests in the laboratory to estimate the effects of copper. They tested this with enclosures in a pond, dosed at two levels of copper, plus a control enclosure. The lab test correctly predicted that *Daphnia* would do well in the low concentration of copper, nor were macroinvertebrates affected. The lab tests did not, however, predict that colonies of alga would decline, as did rotifers and various kinds of copepods (crustaceans).

"We conclude that community responses are complex and cannot be reliably predicted with single-species toxicity tests." Despite the agreement on copper concentrations suitable for *Daphnia*, this experiment must be taken as evidence of disagreement, with the algae and some micro-animals in the mesocosm more sensitive than the single-species lab tests on an animal.

Mount, D.I. and T.J. Norberg-King (eds) 1985. Validity of effluent and ambient toxicity tests for predicting biological impact, Scippo Creek, Circleville, Ohio. United States Environmental Protection Agency, Environmental Research Laboratory, and Permits Division, Washington, D.C. EPA/600/3-85/044.
This is a small sunfish/bass creek flowing through an agricultural area, receiving an effluent from a chemical resin plant. The effluent proved to be non-toxic and no biological effect was found in the stream except for a small area around the outfall where the substrate had been changed.

Mount, D.I. and T.J. Norberg-King (eds) 1986. Validity of effluent and ambient toxicity tests for predicting biological impact, Kanawha River, Charleston, West Virginia. U.S. Environmental Protection Agency, Environ. Res. Lab., Duluth, Minn., and Permits Division, Washington, D.C. Rept EPA/600/3-86/006.
They studied 125 km of this navigable river, which has many municipal and industrial discharges.

Toxicity tests were carried out on some of the effluents but were not a planned part of the study; they failed to predict toxicity that was present in the river. Ambient sublethal toxicity tests with fathead minnows and *Ceriodaphnia* were said to show 60% to 100% correct prediction of instream community effects, depending on the severity of effect compared. There was a high correlation between ambient toxicity to *Ceriodaphnia* and effects on number of species of zooplankton, over the entire length of river. Periphyton and benthic invertebrates were surveyed as well as the plankton. The ambient toxicity tests underestimated the effects on macroinvertebrates.

Mount, D.I., T.J. Norberg-King and A.E. Steen (eds) 1986a. Validity of effluent and ambient toxicity tests for predicting biological impact, Naugatuck River, Waterbury, Connecticut. United States Environmental Protection Agency, Environmental Research Laboratory, Duluth, Minn., and Permits Division, Washington, D.C. EPA/600/8-86/005.
The toxicity tests with effluent were not suitable for comparison with instream effects, having been designed for another purpose. Ambient toxicity tests were compared, with a claim of up to 85% correct prediction at four levels of impairment (% change from a "normalized" value established as best performance found among the sampling stations).

Mount, D.I., A.E. Steen and T.J. Norberg-King (eds) 1985. Validity of effluent and ambient toxicity testing for predicting biological impact on Five Mile Creek, Birmingham, Alabama. U.S. Environmental Protection Agency, Environ. Res. Lab., Duluth, Minnesota, and Permits Division, Washington, D.C. Rept EPA/600/8-85/015.
Biological study included periphyton, macroinvertebrates, plankton, and fish. Ambient toxicity tests were done but the main comparison was done with effluent toxicity to fathead minnows and *Ceriodaphnia*. Calculations of dilution led to prediction of effects at three stations below coke plants and a municipal waste treatment plant. Communities at those stations were affected, with numbers of species reduced by one half or more. "The data from this study clearly indicate the utility of effluent toxicity tests." The authors claim that the single-species tests and community surveys showed agreement in 85% of the cases, but most of this was correctly predicting no impact. Another conclusion was that no single test species or community group is suitable for revealing the impact at every station.

Mount, D.I., A.E. Steen and T.J. Norberg-King (eds) 1986b. Validity of effluent and ambient toxicity tests for predicting biological impact, Back River, Baltimore Harbor, Maryland. U.S. Environmental Protection Agency, Environ. Res. Lab., Duluth, Minn. and Permits Division, Washington D.C. Rept EPA/600/8-86/001.
This was an estuary near Baltimore. They used the sublethal tests with freshwater daphnids and fathead minnows to assess the effluent and also for some ambient tests with parallel tests of salinity effects. Microtox was also used. Field studies assessed zooplankton, macrobenthos and fish. There were too few species in the estuary to establish pollutional changes. Toxicity tests with effluent and surface water agreed.

Mount, D.I., A.E. Steen and T.J. Norberg-King (eds) 1986c. Validity of ambient toxicity tests for predicting biological impact, Ohio River, near Wheeling, West Virginia. U.S. Environmental Protection Agency, Environ. Res. Lab., Duluth, Minn., and Permits Division, Washington, D.C. Rept EPA/600/3-85/071.
They studied 12 km of this large navigable river, a section that receives many effluents. Instream surveys included plankton, periphyton and benthic macroinvertebrates. No toxicity tests were done on effluent. Sublethal toxicity tests on river water samples used fathead minnows and *Ceriodaphnia*. There

was a general correspondence of ambient toxicity to *Ceriodaphnia*, with small effects on communities, and variation in toxicity paralleled the number of species of resident macroinvertebrates. The authors claim from 63% to 100% agreement from single-species toxicity to community effects.

Mount, D.I., and C.E. Stephan 1967. A method for establishing acceptable toxicant limits for fish – malathion and the butoxyethanol ester of 2,4 -D. Trans. Amer. Fish. Soc. 96: 185 -193.

Mount, D.I., N.A. Thomas, T.J. Norberg, M.T. Barbour, T.H. Roush and W.F. Brandes 1984. Effluent and ambient toxicity testing and instream community response on the Ottawa River, Lima, Ohio. United States Environmental Protection Agency, Permits Division, Washington, D.C., and Environmental Research Laboratory, Duluth, Minn. EPA -600/3-84-080.
This is representative of the eight "validation studies" by U.S. EPA. Point-sources were a municipal treatment plant and a refinery. They studied toxicity of the effluent with 7-day growth of larval fathead minnows and 7-day reproduction of *Ceriodaphnia*, and also included high concentrations to measure acute lethality. They also ran tests for acute lethality of effluent with representatives of eight families of resident fish. They tested with *Ceriodaphnia* for sublethal toxicity of the receiving water at selected locations ("ambient" toxicity tests).

The field survey included quantitative assessment of the periphyton, macroinvertebrates, and fish. Fish were also exposed in cages to assess mortality (species not identified). Nighttime studies of the drift of invertebrates were done in a few locations. Hydrographic work was also done to identify plume concentrations at the discharges, and basic chemical characterization was also done.

The instream impact ended at a point where the ambient toxicity tests also showed no sublethal effect. "A correlation was established between ambient toxicity, effluent toxicity and biological impact which suggests that effluent and ambient toxicity tests are accurate predictors of receiving water impact."

Moverley, J.H., D.A. Ritz and C. Garland 1995. Development and testing of a meiobenthic mesocosm system for ecotoxicological experiments. National Pulp Mills Research Program Tech. Rept No. 14. CSIRO, Canberra, Australia. 117 p.
An attempt to develop a mesocosm for pollution studies, based on small sediment-living marine animals. The authors admit that results are too variable to be very useful at present. "However, further research needs to be done ... before mesocosms can be fully developed as routine monitoring tools. At this stage, the ... mesocosm ... recommended by this study should be considered only as a basis for further work."

Munkittrick, K.R., G.J. Van Der Kraak, M.E. McMaster, C.B. Portt, M.R. van den Heuvel and M.R. Servos 1994. Survey of receiving-water environmental impacts associated with discharges from pulp mills. 2. Gonad size, liver size, hepatic EROD activity and plasma sex steroid levels in white sucker. Environ. Toxicol. Chem. 13: 1089-1101.
An example of using physiological "biomarkers" in fish as a warning of potential deleterious effects. The abnormalities measured at many mills agreed in general with deleterious effects seen in biological surveys of invertebrates. They did not, however, show up as whole-body effects on wild fish (growth etc.); it was not known whether changes occurred at the population level.

Norberg, T.J., and D.I. Mount 1985. A new fathead minnow *(Pimephales promelas)* subchronic toxicity test. Environ. Toxicol. Chem. 4: 711 -718.

Norberg-King, T.J. and D.I. Mount (eds) 1986. Validity of effluent and ambient toxicity tests for predicting biological impact, Skeleton Creek, Enid, Oklahoma. U.S. Environmental Protection Agency, Environ. Res. Lab., Duluth, Minn. Rept EPA/600/8-86/002.
This stream in an agricultural area received discharges from an oil refinery, a fertilizer plant and a municipal waste treatment plant. Instream surveys covered plankton, macroinvertebrates and fish. Prediction from the sublethal effluent tests with fathead minnows and *Ceriodaphnia* identified the most affected station. The authors claim 87.5% agreement between "lab" and field results. "The data from this study clearly indicate the utility of effluent and ambient toxicity tests for predicting instream effects."

Norberg-King, T.J., D.I. Mount, J.R. Amato, D.A. Jensen and J.A. Thompson 1991. Toxicity identification evaluation: characterization of chronically toxic effluents, phase 1. United States Environmental Protection Agency, Office of Research and Devt, Natl Effluent Toxicity Assessment Center, Duluth, Minn. EPA -600/6-91/005.

OECD (Organization for Economic Co-operation and Development) 1992. Report of the OECD workshop on the extrapolation of laboratory aquatic toxicity data to the real environment. OECD Environment Monographs No. 59. Paris, France. OCDE/GD(92)169.
Recommendations from a workshop. Lots of good principles and recommendations. No specific examples of comparisons/validation that are useful in the present review.

Okkerman, P.C., E.J. Plassche, C.J. Roghair and J.H. Canton 1990. Validation of some extrapolation methods with toxicity data derived from multiple species experiments. Presented at the OECD workshop on ecotoxicological effects assessment, Arlington, Va. National Institute of Public Health and Environmental Protection, Bilthoven, The Netherlands. 14 p. + app.
This is the predecessor of Emans et al. 1992, use that one, which is more complete.

Perrin, C.J., B. Wilkes and J.S. Richardson 1992. Stream periphyton and benthic insect responses to additions of treated acid mine drainage in a continuous-flow on-site mesocosm. Environ. Toxicol. Chem. 11: 1513-1525.
This does not have a comparison of effluent tests and field surveys, but it is an example of the use of mesocosms with a mine discharge in British Columbia. The authors recommend study of the algal and insect communities developing in on-site, flow-through troughs as a way of assessing treated acid mine drainage. They found no effect of 10% concentrations.

Persoone, G. and C.R. Janssen 1992. Field validation of predictions based on laboratory toxicity tests. Proc. European Workshop on Fresh Water Field Tests, Potsdam, Germany, June 25-26, 1992, 32 p.
A general review, too extensive to summarize here. Some useful examples and useful synopses of important papers by other authors. Parts of this review are quoted at various places in the present document.

"... generally spoken, the conclusions ... are confirmed in the large majority of cases, namely that experimental (or calculated) NOECs from laboratory [single-species] tests relate "relatively well with the highest toxicant levels which do not have effects on populations in the field." They go on to point out that the prediction is improved as more species are included in the laboratory tests, and especially if the range of species is chosen well.

Persoone, G., A. Van de Vel, M. Van Steertegem and B. De Nayer 1989. Predictive value of laboratory tests with aquatic invertebrates: influence of experimental conditions. Aquat. Toxicol. 14: 149-166.
This paper shows the importance of testing under the ancillary conditions prevailing in the surface water of interest, i.e. the best procedure is to use upstream/clean water as the dilution water in tests. Although the paper deals with tests of acute lethality, it provides a well-documented example of the effects of modifying factors. Three hundred tests of acute lethality were done on a few reference toxicants, at various levels of environmental variables (temperature and hardness or salinity). Results are plotted up

as neat histograms. The variation in toxicity with species and modifying condition, ranged from 2.5 to 100.]

Pontasch, K.W., B.R. Niederlehner and J. Cairns Jr. 1989. Comparisons of single-species, microcosm and field responses to a complex effluent. Environ. Toxicol. Chem. 8: 521 -532.
A complex study with: (1) laboratory tests of acute lethality to *Ceriodaphnia*, *Daphnia magna*, and fathead minnows; (2) laboratory test of *Ceriodaphnia* reproduction; (3) laboratory dosing of microcosms of macroinvertebrates; (4) laboratory microcosms of protozoa colonies; (5) field (*in situ*) exposures of microcosms of protozoa; and (6) field samples of macroinvertebrates from the polluted stream which was a pristine waterway polluted with a single complex effluent (type unspecified). The results are synopsized below, with some interpretation by the present author to provide actual estimates of NOEC, LOEC, and TOEC (threshold-observed-effect concentration, the geometric mean of LOEC and NOEC). These numerical estimates replace some of the general discussion given by the authors of the paper.

Type of test and response	LOEC	NOEC	TOEC
Laboratory			
(2) *Ceriodaphnia* reproduction, 7-d	3 %	1 %	1.7 %
(3) Microcosms, macroinvertebrates	1 %	0.1 % *	0.32 %
(4) Microcosms, protozoa	1%	0.1 %	0.32 %
Field			
(5) Microcosms, protozoa	14.1 %	4.1 %	7.6 % **
(6) Field samples, macroinvertebrates	3.5 %	1.1 % *	2.0 % **

* A very slight effect was seen, but considered by J.B.S. to be not ecologically significant for purposes of the present analysis.

** Calculation based on interpretation of NOEC and LOEC by J.B.S., supplementing the discussion by the authors, who designated the various significant effects but did not actually name overall values for NOEC and LOEC.

Of most interest to the present chapter is comparison of (1) the single-species laboratory test, with (6) the field study of resident invertebrates. The laboratory TOEC of 1.7 % is very close to the field value of 2.0 % effluent.

The field microcosm (5) is of the same order of magnitude as the field survey (6). It is of interest that the lab microcosms (3) and (4) are an order of magnitude more sensitive than the field microcosm (5) and survey of resident fauna (6). The authors comment that this might be because of faster degradation of some components of the effluent in the stream, compared to the laboratory.

The authors were apparently most interested in discussing the merits of microcosms as a useful test method; they give generally subjective or detailed interpretive conclusions, rather than a bare numerical comparison as attempted in the table above. "Microcosm responses corresponded well with observed effects in the field. The microcosms correctly predicted which indigenous organisms would be lost and which would be stimulated at various ... concentrations of the effluent."

Robinson, R.D., J.H. Carey, K.R. Solomon, I.R. Smith, M.R. Servos and K.R. Munkittrick 1994. Survey of receiving-water environmental impacts associated with discharges from pulp mills. 1. Mill characteristics, receiving-water chemical profiles and lab toxicity tests. Environ. Toxicol. Chem. 13: 1075- 1088.
They studied 11 Canadian pulp and paper mills. They did not do sublethal toxicity tests on effluents, but they did do them (*Ceriodaphnia* and fathead minnow larvae) in the receiving water upstream and downstream of the mills. They reviewed available information on communities, mostly studies of benthic invertebrates, and made general comparisons (non-statistical).

"Fathead minnow and *Ceriodaphnia* tests were generally correlated with historical data on benthic macroinvertebrate community responses."
"[I]t appears that the fathead minnow and *Ceriodaphnia* tests conducted in the present study were somewhat predictive of the degree of impact on the benthic macroinvertebrate community. This predictive ability exists, despite the fact that benthic community effects caused by pulp mill effluent are usually attributed to BOD and solids loadings, with the resulting effects on substrate physical and chemical ... characteristics."
"It is difficult to draw conclusions regarding the ability ... to predict fish community responses. The current survey was not designed to evaluate changes in fish species diversity and abundance, ... and previous studies of this nature are limited."

They point out that the sublethal toxicity tests did not correlated with the physiological/biochemical changes (enzyme activity, sex steroid levels, gonad size, etc.) found within fish collected downstream of the mills. The meaning of this is obscure because it was not clear that the within-fish changes were associated with deleterious effects on whole-organism performance or populations of fish. In other words, the physiological/biochemical findings might be the ones that are out of step with other findings.

Rosenzweig, M.S. and A.L. Buikema, Jr. 1994. Phytoplankton colonization and seasonal succession in new experimental ponds. Environ. Toxicol. Chem. 13: 599-605.
This paper relates to mesocosms as an experimental method of validating predictive tests. They built twelve ponds, each 20 m square, and intended to be replicates. Specifications followed specifications of U.S. EPA for cosm toxicity tests. The ponds were allowed to colonize themselves and the developing communities were followed for a year.

"Similar successional patterns in all 12 ponds occurred; however, the community structure between ponds was not similar at any time. ... after one year they were not mature enough for use as replicated test systems. ... mesocosms need to be managed to produce replicate experimental units."

Samoiloff, M. 1990. The nematode toxicity assay using *Panagrellus redivivus*. Toxicity Assessment: An Internat. J. 5: 309-318.

Sanders, W.M. 1985. Field validation. P. 601 -618 *in:* G.M. Rand and S.E. Petrocelli (eds). Fundamentals of aquatic toxicology. Methods and applications. Hemisphere Pub. Corp., Washington, D.C.
This review gives excellent advice on how to *design* validation programs. It is quite general, and applies to a diversity of programs involving chemistry as well as biology, but has principles that apply to validation of toxicity tests on effluents. It should be read when drawing up plans for validation. One of the interesting pieces of advice is that if economics or resources do not allow an adequate project,

abandon it rather then producing an invalid attempt at validation.

Scroggins, R.P. 1986. In-plant toxicity balances for a bleached kraft pulp mill. Pulp and Paper Canada 87 (9): T344-348.

Slooff, W., J.H. Canton and J.L.M. Hermens 1983. Comparison of the susceptibility of 22 freshwater species to 15 chemical compounds. I. (Sub)acute toxicity tests. Aquatic Toxicol. 4: 113-128.

Slooff, W., J.A.M. van Oers, and D. de Zwart. 1986. Margins of uncertainty in ecotoxicological hazard assessment. Environ. Toxicol. Chem. 5: 841-852.
This massive analysis of data focuses on acute tests, so most of it is of reduced interest for the present review. However, parts deal with sublethal effects, and there is considerable relevance of one topic, the third one listed below, which deals with single-species versus multi-species tests. There were three main conclusions.

First, there were great differences between species, whether they belonged to the same group or different groups.

Second, there was good prediction of chronic toxicity from acute toxicity for the same species. They developed a general predictive relationship based on sets of data for 164 chemicals; it predicts the sublethal NOEC from the acute LC50 or EC50.

$\mathrm{NOEC} = -1.28 = 0.95 \log[\mathrm{L(E)C50}]$ This has a correlation of $r = 0.89$.

By and large, this would signify an acute-chronic ratio of about 20, but varying from about 13 for very toxic substances, to about 30 for mildly-toxic substances.

Third, multi-species or "ecosystem testing does not lead to results that are dramatically different from those obtained with single-species tests". They provide predictive relationships for this also. One relationship (not repeated here) estimates the ecosystem no-sublethal-effect level (NOEC) from the acute effective concentration. Another relationship is very relevant here since it relates sublethal field assessment to sublethal tests in the laboratory. The model-ecosystem NOEC may be predicted from a single-species NOEC as follows.

$\mathrm{NOEC(ecosystems)} = 0.63 + 0.85 \log[\mathrm{NOEC(species)}]$ For this, $r = 0.85$.

The conclusion on this topic indicates that single-species effluent tests could be indicative of effects in the communities of receiving waters.

Snell, T.W. and B.D. Moffat 1992. A 2-d life-cycle test with the rotifer *Brachionus calyciflorus*. Environ. Toxicol. Chem. 11: 1249-=1257.
A very promising test. The organisms are obtained as cysts (commercial source) and so there is no culture involved. The exposure represents 1.3 generations and reproduction is assessed by the number of organisms. It is said to take 70% fewer person-hours than a test with *Ceriodaphnia*, while preliminary comparisons indicate similar sensitivity.

Sprague, J.B. 1985. Factors that modify toxicity. P. 124 -163 *in:* G.M. Rand and S.E. Petrocelli (eds). Fundamentals of aquatic toxicology. Methods and applications. Hemisphere Pub. Corp., Washington, D.C.
Many examples and discussion of how the ancillary factors (oxygen, hardness, pH, etc.) can affect the toxicity measured. Provides reasons why tests should be done under realistic conditions similar to the conditions in the waterbody of interest.

Sprague, J.B. 1986. A simple in-stream test of laboratory findings that NTA protects fish and invertebrates against copper and zinc. P. 213-223 *in: Community toxicity testing* . (ed. J. Cairns Jr.). American Society for Testing and Materials, Spec. Tech. Pub. No. 920.
This validation was connected with metal-mining pollution in northeastern New Brunswick. The chelating agent nitrilotriacetic acid (NTA) was considered as a temporary remedy for surges of pollution until control measures were operative. Laboratory tests had shown that NTA protected fish from toxicity of zinc and copper. This test in a small stream showed that the predicted protection also applied in nature. In a section of stream exposed for four days to 2.4 to 6 times the lethal level of the metals, simultaneous dosing with equimolar NTA resulted in no apparent harm to caged fish, and normal behaviour of wild eels. In an unprotected section of stream, fish died within 4.5 hours, and dead wild eels were also found. The field results were as predicted from the lab tests. For invertebrates, laboratory tests had not been done, so no predictive validation was possible, but most invertebrates proved to be more tolerant than expected from other field observations.

Sprague, J.B. 1991. Environmentally desirable approaches for regulating effluents from pulp mills. Water Sci. Technol. 3/4: 361-271.
The general content of this paper is contained in section 3 of the present chapter.

Sprague, J.B., P.F. Elson and R.L. Saunders 1965. Sublethal copper -zinc pollution in a salmon river – a field and laboratory study. Internat. J. Air Water Pollution 9: 531 -543.
This deals with avoidance reactions of fish to pollutants, a special topic but part of community existence in polluted rivers. Laboratory tests with young salmon showed that they avoided copper-zinc mixtures at a threshold (LOEC) of 0.02 toxic units. Adult wild salmon migrating upstream in a New Brunswick river appeared to show avoidance reactions (return downstream through a counting fence) at metal concentrations of 0.35 to 0.43 toxic units. The wild reaction was 17 to 20 times higher than the laboratory prediction. It was concluded that much of this difference would be the "urge" of adult salmon to move upstream, whereas in the laboratory tests there was no over-riding motivation to choose one side or another of the experimental tank, except for the presence of added metal on one side.

SRC 1995a [Saskatchewan Research Council]. *Lemna minor* toxicity test. SRC Water Quality Lab., Regina. Standard Operating Procedure. SOP 199, 10 p.

SRC 1995b [Saskatchewan Research Council]. Phytoplankton microplate growth test using fluorescence as the endpoint. SRC Water Quality Lab., Regina. Standard Operating Procedure. SOP 204, 23 p.

Stephan, C.E., D.I. Mount, D.J. Hansen, J.H. Gentile, G.A. Chapman and W.A. Brungs 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. U.S. Environmental Protection Agency, Office of Research and Development, Washington, D.C.. [Available as NTIS PB 85-227049].

Times Ltd. 1994. Review of the WET test program for HMC outfall 001 NPDES permit no. SD-0000043. Times Limited, 82777 Cottonwood Road, Bozeman, Montana 59715, U.S.A. 19 p. + figure, tables, and appendices.
WET is used in U.S.A. to signify whole effluent toxicity. The toxicity data are provided, and the report on biological survey is given as an appendix. Much of the report details the treatment methods and efforts to find toxic components of the wastewater.

Thurston, R.V., T.A. Gilfoil, E.L. Meyn, R.K. Zajdel, T.I. Aoki and G.D. Veith 1985. Comparative toxicity of ten organic chemicals to ten common aquatic species. Water Res. 9: 1145 -1155.

Warren, C.E. 1971. Biology and water pollution control. W.B. Saunders Co., Philadelphia, Pa.

Warren, C.E. and G.E. Davis 1971. Laboratory stream research: objectives, possibilities, and constraints. Annual Rev. Ecol. Systematics 2: 111 -144.

Zischke, J.A., J.W. Arthur, R.O. Hermanutz, S.F. Hedtke, and J.C. Helgen 1983a. Effects of pentachlorophenol on invertebrates and fish in outdoor experimental channels. Aquatic Toxicol. 7: 37-58.
A good example of the use of artificial outdoor streams, although there were no replicates except for testing twice. They did sublethal laboratory tests to estimate the "safe" concentration of 48 μg/L, and it was close to being correct as determined by thorough 12-week stream tests in two years. The presumed "safe" concentration reduced fish growth and raised larval drift. There were, however, no effects on macroinvertebrate density or species, or on survival and reproduction of fish.

Zischke, J.A., J.W. Arthur, K.J. Nordlie, R.O. Hermanutz, D.A. Standen and T.P. Henry 1983b. Acidification effects on macroinvertebrates and fathead minnows *(Pimephales promelas)* in outdoor experimental channels. Water Res. 17: 47-63.
Good "field work". Diverse findings on species sensitivity as well as community characteristics. There is a large amount of information in the literature on tolerance of pH by aquatic organisms, which could be checked against this excellent study. No specific laboratory work was done in this study.

Appendix. List of 27 reports retrieved from the AQUAMIN data-base

These were reports which included both toxicity in the laboratory, and study of populations or communities of macroinvertebrates.

The number of the report in the AQUAMIN bibliography is given first. Then selected information from the AQUAMIN "key words" and "remarks" is given. [Remarks in square brackets are those of the present author.]

# 6 Lethal tests. [Lethal tests are not useful in the present report, only sublethal ones.]
# 22 Toxicity tests with sediment.
# 162 Only a summary, see # 673.
# 172 Lethal tests.
# 173 Reduced diversity of benthos. [Type of toxicity test not indicated.]
# 204 One salmon toxicity test. [Apparently lethal.]
# 222 One lethal test but no LC50. Limited sampling, interpretation difficult.
# 226 **Chronic test** [no such test in report]. Also lethal test. Rainbow trout.
# 243 Annual environmental assessment. Data in # 249. [Lethal tests?]
# 246 As # 243, but data in # 242.
# 247 As # 243, but data in # 248.
# 271 **Sublethal** and lethal tests. Fathead minnows, salmon. Detailed assessment.
# 284 [No remarks. Sublethal tests?]
# 289 **Sublethal** and lethal tests, rainbow trout and Daphnia.
# 396 Microtox, Daphnia, trout [acute/lethal?]. No methods, technically not a very good report.
# 436 Lethal. Summary of monitoring, results only, consult original reports.
# 490 Overview, conclusions only, document appears incomplete.
# 495 **Chronic** [sublethal?] and acute. Good assessment of aquatic effects.
# 497 Excellent study of impact on benthos, trout growth. [Lethal tests?]
# 546 Lethal tests. Appendix B [toxicity tests] is missing. Same as # 618.
# 584 Marine. [Lethal tests?]
# 607 Microtox, Mutatox tests. Ten sediment toxicity tests. Follows # 606.
# 618 Same as # 546. Appendix B missing [toxicity tests in lab ofJ.B. Sprague].
# 642 Lethal tests. Study done after a discharge. Methods, results elsewhere.
# 666 Baseline, primarily a data report. Toxicity tests [lethal?] on bluegill, rainbow trout.
# 672 Lethal tests.
# 724 Two 30-d *in situ* toxicity tests. Vol. 2, with data not provided.

Chapter 3.

Planning a field validation program.

Table of contents. Chapter 3

1 Introduction 96

- 1.1 Selecting the degree of scientific stringency 96
- 1.2 Generality versus specificity in this chapter ... and a sermon 97

2 General advice 98

3 Steps in a scientific field validation 102

- Phase 1. Verify the toxicity test(s) 102
- Phase 2. Select the sites 103
- Phase 3. Survey existing records 104
- Phase 4. Initial screening and sampling program 104
- Phase 5. Preliminary modelling/design/analysis 105
- Phase 6. Design of detailed validation program 105
  - *(Item 6a) Scientific method* 105
  - *(Item 6b) Sensitivity analysis* 106
  - *(Item 6c) Logistical support plan* 106
  - *(Item 6d) Cost-benefit analysis* 107
- Phase 7. Carry out the program 107

4 Sources of information on planning field validation 108

References 108

1 Introduction

In the contract governing this report, the field program is described in the following words[1] .

> "Evaluation of the ability of the final short list of chronic (*sic.*) toxicity tests to measure the contribution of various discharges (point and non-point source) to observed biological effects in the aquatic receiving environment, as component of the field studies 1996/97 and 1997/98." The evaluation was planned for three sites in each of the time-periods.

The contract requires Sprague Associates Ltd. to "recommend the design of the field program", and the requirement is being met by this chapter. Comments and advice given here are intended to fit the field program given above.

1.1 Selecting the degree of scientific stringency

Various degrees of scientific sophistication could be applied in the field program of AETE. We might distinguish two important categories.

> In *field evaluation* the degree of correlation could be determined, between (a) the results of single-species toxicity tests and (b) measurements of the quality of the receiving community.
>
> In *scientific field validation* there would be a higher degree of rigour, i.e. a formal scientific test of whether or not the toxicity tests predicted the effect on the community.

Sanders (1985) provides the terminology above from definitions of the U.S. Environmental Protection Agency (EPA), notably the following one.

> *Field validation is the process of comparing the overall result or output of a method, toxicity test, or model with data obtained from real-world environments.*

From accompanying wording in Sanders' chapter, it is clear that the validation process requires formal scientific testing, with acceptability criteria set beforehand, so that the toxicity test would either agree or disagree with the field results. In other words it would be valid or invalid for predicting conditions in the field. Obviously that is an explicit, cut-and-dried process, not merely the establishment of some degree of correlation[2] .

---

[1] Item 3 of the contract deals with "Background". It lists the three major steps of AETE's sublethal toxicity program.

[2] Sanders (1985) pointed out that relatively few true field validations had been done in the preceding two decades, although many tests had been evaluated for degree of correspondence with real-world data. The same comment on few scientifically rigorous comparisons among many general comparisons was made in chapter 2, which dealt specifically with validation.

Similarly, Cairns et al. (1988) review validations in water pollution biology, and distinguish the degrees of scientific sophistication. For the simpler correlation approach (= field evaluation, above), they use the term *preliminary validation*. For testing prediction in a formal manner they use the term *predictive validation*, and make it clear that this requires the sequence (a) toxicity tests, then (b) prediction to the community, and finally (c) measurements in the field to assess the accuracy of the prediction. Cairns (1986, 1988) deals with the topic in a generalized way that is less focused on the interest here.

The AETE Toxicity Testing Sub-group (hereafter called the "committee") can select the degree of scientific sophistication that it wishes to have. The choice will involve priorities and compromises on objectives, dependability of results, timing, and resources available. Only the committee can decide on these priorities and hence on the scientific stringency.

This chapter cannot hope to describe a whole series of alternative programs at various levels of scientific elegance. The simplest way to present information to the AETE committee is to describe the steps in *scientific field validation*, i.e. the high level of scientific investigation. That will provide a comprehensive description. For less rigorous programs, it will be fairly obvious to the committee where the program can back off to a less demanding design. A less sophisticated program might be desirable, for example, in order to release resources for study of additional minesites or alternative topics.

The subject of this chapter is, accordingly, *scientific field validation* of sublethal toxicity tests. The main objective of the field program might be rephrased for purposes of this chapter.

> *To determine whether selected sublethal toxicity tests on mine wastewater predict biological effects of the wastewater in the receiving waterbody.*

1.2 Generality versus specificity in this chapter ... and a sermon

Most of the advice in this chapter is of a general nature, emphasizing planning by use of the scientific method. All the advice applies directly to the AETE field program, although much of it would apply to field programs with somewhat different goals. In other words, the advice given here could be relevant, even if AETE modified the purposes of its field program.

It would have been counter-productive to attempt here, a specific design for a field program (e.g. "take 3 samples of benthic organisms at each of the 5 sampling stations, every two months"). Objectives and methods which might seem reasonable to the author, would be unlikely to fit the concepts and objectives of the AETE committee.

On the other hand, I am confident that the general approach outlined here for scientific planning, could lead to an economical and rewarding program, providing useful and dependable answers. Probably there are few of us, certainly not this author, who could say that we had conscientiously followed the good principles of the scientific method at every point in our technical and scientific careers. Sometimes it seems necessary to deviate from them for one reason or another. Nevertheless, following the scientific principles leads to sound and highly-productive experiments.

As a biologist, perhaps I can be allowed to sermonize that the AETE committee should check that these principles are evident in proposals for the field program. We biologists often tend to fly by the seat of our pants. We tend to rush into the field and collect samples, or set up a lab and do toxicity tests, process a bunch of results, then perhaps take the numbers to a statistician to see if anything shows up. I have heard biologists say that they were going to launch a field study and "collect all the information possible". Presumably whether it was relevant or not.

AETE has an opportunity to carry out a program that will be useful and informative. Alternatively, there is an opportunity to carry out a fair amount of busy work. The latter might add little to our scientific capability for monitoring water quality in an economical fashion. Selecting a good field program will take a little thought on goals, methods, and available resources.

2 General advice

(1) Use the scientific method.
Some scientists do this. They claim that the method is powerful and productive. It requires asking oneself what is the purpose of the work before setting out to do it. Beyond that, one must formulate an hypothesis, *design the simplest possible experiment* to test the hypothesis, decide the criteria for rejection, carry out the work, test statistically, then draw conclusions.

(2) Take advantage of the scientific literature.
Many topics of interest have already been studied widely. It can be faster and cheaper to answer a given question by taking advantage of results in the literature, rather than setting up a new experiment. It can also be much more definitive to use the literature, since there are often dozens or hundreds of publications that illuminate the topic. In comparison, a new experiment might produce only a small amount of information. A footnote lists four examples of topics that might be expeditiously assessed from the literature[3] .

---

[3] Artificial substrates for sampling bottom-living organisms are discussed in many papers, and at least one book (Cairns 1982). Values, drawbacks, and validity of these methods are covered.

Many publications going back several decades deal with methods of sampling aquatic invertebrates and processing the samples. A superb review of methods for biological surveys of freshwater pollution was published by Hellawell (1978) and still provides good orientation.

Beyond that there is literature on advantageous approaches to taxonomic identification and indices of community structure. A current frontier is the use of biotic indices to categorize degree of pollution; they are specially designed to reflect *water quality* as opposed to other types of habitat degradation. Biotic indices of water pollution enjoyed vigorous development in Europe

(3) Do not try to do too many things.
Since funding will presumably be fixed, it would be undesirable to spread the resources thinly and study a number of topics in an insubstantial fashion. It would be more profitable to pick the most important question remaining after an initial scan of literature, then to tackle that question with a program that was powerful enough to yield an answer.

One apparent question for this program was stated at the beginning of the chapter, i.e. validation of effluent toxicity tests as predictors of effect in the receiving community. General confirmation of the predictive ability of effluent tests is found in the literature (chapter 2). The AETE committee might consider that the evidence in the literature is convincing, and that no validation project was required under AETE sponsorship. Alternatively, it might be desired to design a specific program of *scientific validation* for the Canadian metal-mining industry.

After assigning the necessary resources for answering a primary question or questions, any remaining resources might be allocated to a selected secondary question or questions. Examples of data-gathering that would not contribute to answering the main question, are given in the preceding and following items, but other examples could be given[4].

(Washington 1984). They became streamlined, easy to use, and robust in assessing many kinds of pollution. New indices are currently being used in the U.S.A., and Karr (1993) provides an introduction to the concepts and literature. The most advanced use of biotic classification is proposed for Britain, where indices based on aquatic invertebrates are used to define the health of the waterbody at a given location. The biotic index is used for monitoring conditions, and there are plans to extend its use, as a graded national water quality guideline (NRA 1991). The index would assign classifications to sections of rivers, and a selected value would be used to set a numerical quality objective in a given location. In case of disagreement of classification by chemical and biological criteria, there is a procedure for "biological over-ride". These extensive developments would be best evaluated through an initial literature review; some aspects might be useful for monitoring effects of metal mining, or a particular facet might deserve field testing for use in Canada.

There is a diverse literature on metallothioneins (MT) in fish; a decade ago there was an astute program related to mining in British Columbia (e.g. Roch and McCarter 1984). It is clear that the presence and amount of MT represents a very complex topic, with no simple role for MT as an index of degree of pollution. Similarly, a large body of literature deals with metal concentrations in fish and other organisms. One of many complicating factors is that fish can deal very well with many common metals, excreting them readily, so the amount remaining in their bodies has no simple relation to exposure. An amazingly large body of information on metals in marine invertebrates proves mainly that accumulation of metals in aquatic animals does not favour simplistic interpretation (e.g. a thousand-paper review, Sprague 1986).

[4] It might be difficult to find scientific justification for sediment toxicity tests as part of a *validation* program. Although there is considerable (and welcome) activity in this long-neglected field, that does not mean that sediment tests should be included in the AETE program. If there were a specific need to validate the ability of sediment tests to predict community effects, then a discrete experiment should be designed to accomplish that validation.

(4) Be cautious about devoting funds to physical and chemical measurements
Routine collection of physico-chemical data should not be done simply because it is customary in routine survey work. It might be largely unproductive in a validation program. Physical and chemical measurements in the effluent, receiving water, and sediments*do not* provide proof of a pollutional effect. That must be documented by detrimental changes in a living aquatic community. The chemical findings would only provide evidence that could be used to predict an effect, and in that respect they are predictive in the same way as are toxicity tests on an effluent are predictive[5].

There are, however, good reasons for selected chemical tests in an AETE validation program. (a) Chemical measurements in the effluent could assess fluctuations in toxic components. Measurements might be quick and efficient by chemistry, and correlation with changes in toxicity should be attempted. (b) It might be a goal to see if selected chemical measurements in the effluent were valid predictors of community effects. (c) Physico-chemical measurements would be essential to delineate the effluent plume, and dilutions achieved. That task should be done in an expeditious manner, by rapid and cheap measurements such as conductance.

(5) Within-organism studies are not appropriate for validation
Design of a program to validate effluent toxicity tests should start from the position that nothing below the community level of organization should be measured ("zero-base planning"). Lower-level items would not provide direct validation because they themselves are only predictive and would require validation at the community level.

Accordingly, there would be no apparent scientific reason for including items such as within-organism biochemistry and fecundity in an experiment to validate toxicity tests. Biochemical and physiological "biomarkers" are currently in vogue and show great promise as early-warning indicators. Nevertheless, biomarkers and studies of tissues should not be used in a study to validate effluent tests, unless there were a deliberate experiment to validate them against the biological communities, in addition to validating the toxicity tests done on effluent.

(6) Narrow the biotic surveys
There is a tendency to think that the entire spectrum of organisms in an aquatic community must be studied to appreciate any damage to the community but the opposite is true -- there is a great redundancy of information among the various groups of organisms. The larger invertebrate organisms living on the bottom of a river or lake (benthic macroinvertebrates) are almost always the best indicator organisms for water pollution[6]. Focusing efforts on those

---

[5] For example, guidelines for chemicals in sediments have recently been published in some jurisdictions. There might be a need to measure such chemical concentrations in a particular project of AETE, but there would not appear to be any particular scientific reason to do that in a field program for validating effluent tests. As in the previous footnote, a specific objective of validating chemical measurements in sediments would warrant a separate hypotheses and experiment which compared the health of the community living on and in the sediments.

[6] Often it is felt that fish must be included in a biological assessment of pollution, perhaps because people are interested in fish, but they are poor biological indicators. They move from place to

organisms will get the "biggest bang for the buck" in field surveys of water pollution, except in unusual situations where the benthic habitat is poor for organisms or difficult to sample. In the present program, such locations should be avoided in favour of better geographical sites.

(7) Include many minesites
For any study of pollution at a single location (say a river), there are many problems in assigning cause for an apparent effect that might be found. One of the greatest problems is the "upstream/downstream" nature of the control and experimental stations in a typical riverine study. There are known (although very gradual) downstream changes of flora and fauna in a clean river, and these are confounded with the supposed effect of an effluent. From a statistician's point of view, the clean and polluted stations are not independent. The same general problem of confounding occurs if one attempts to select a control from a nearby similar waterbody. At a given location there could be many influences making the upstream and downstream communities different, such things as depth, substrate, velocity of flow, streambank erosion, or other sources of pollution which were perhaps unknown to the investigator.

There would be a stronger case for attributing changes to the effluent, and not other causes, if there were a potent effect in the river below the discharge, followed by an extensive downstream gradient of concentration and effect, because of downstream dilution by tributaries. However, it might not be easy to find such uncomplicated text-book patterns.

---

place and therefore do not necessarily reflect the pollutional status of the place they are caught. Kaesler et al. (1974) found them to be the worst indicators of water pollution among the various groups of aquatic organisms. Nor were they good indicators of pollution involving sediment deposition (Berkman et al. 1986). In addition fish are sometimes difficult to catch resulting in biased samples. A true assessment of the fish community should include estimates of population size for the various species, and accurate estimates of that can be extremely difficult and expensive. It would be advantageous in this program to avoid study of fish populations, and devote resources to more fruitful topics or to additional locations.

Micro-organisms, whether protozoans or algae, would seem to be poor pollution indicators because of their rapid life-cycle. They can recolonize an area rapidly after a pollution surge, thus failing to reflect the event. However, one group of algae (diatoms) have been declared good pollution indicators (Kaesler et al. 1974). The larger rooted plants are usually not ubiquitous in their distribution, especially in running water, and hence they are also poor indicators.

Benthic macroinvertebrates are excellent biological indicators because the various groups vary in sensitivity providing scope for an index, and because they are relatively fixed in one location. They have a life cycle of months to a year or more and will thus reflect past conditions. Semi-quantitative sampling is usually feasible, and handling and identification are relatively easy.

Economical pollution surveys can be done by identifying and tabulating only part of the biota. Kaesler et al. (1974) compared all kinds of organisms and concluded that aquatic insects had the best overall correlation with information from other groups. Within the insects, "information provided by mayflies alone would probably have been sufficient to assess recovery of the stream". Beetles were also good; the two groups of insects gave most of the information from surveying all insects.

Another remedy is to study many locations. Conclusions become more convincing if all the locations show a relationship of community change to the effluent discharge, and also show a degree of change that is related to the toxic concentration. Over many locations, other factors such as current velocity might cancel themselves out because they were not consistent with the main pattern of effects. Given results for only one location or a few locations, it would be impossible to rule out the extraneous factors; given many locations, it might be feasible.

Such basic questions of design are obviously deeply involved with the methods selected for statistical analysis of the results. Neither the overall design or the method of analysis should be capriciously chosen by itself, they must develop together. Clearly, it is essential to involve a statistician, starting with the beginning of the program design[7]. Publications of Green (1979, 1990) have statistical advice on biological surveys which might assist the non-statistician.

(8) Validation studies in lakes can be expected to be difficult
There are several potential problems. Currents driven by wind and other causes could be variable, and the discharge plume might move around. Thus a relatively constant gradient of concentration might not be available as a base for community studies, as would be the case in rivers. The whole of a small lake might be at essentially the same concentration of wastewater, preventing a study of graded effects. Depth of water could confound the biological sampling, if benthic organisms were influenced by differences in temperature and oxygen at different depths. Type and degree of sediment deposit would also affect the biological community and might be unrelated to wastewater. If validation could be accomplished in rivers of suitable size, that would be a good and satisfactory accomplishment for this program.

3 Steps in a scientific field validation

The following phases or stages of an investigation are numbered sequentially and for the most part the fall into the order logically. Some phases, however, would overlap in timing.

Phase 1. Verify the toxicity test(s)

This is not a problem in the present program. Verification simply means confirmation that the test operates as intended. The current AETE program of assessing sublethal tests in the laboratory will provide verification. Most of the tests being considered have already been verified by wide experience elsewhere and by numerous published studies of their performance.

---

[7] In particular, the number of minesites, be it 10, 20, or 30, should follow advice of the statistician after it is known which tests are to be done and what is their variation. Potential breadth of the study would presumably decrease as number of locations increased, because of limits on total effort.

When carrying out the toxicity tests in the field validation program, it is of utmost importance that dilution water for the tests should be taken from "upstream" in the waterbody being studied. That will integrate any modifying effects of the water on toxicity, making the tests as relevant as possible to the community studies[8].

Phase 2. Select the sites

Site selection (and indeed the whole design process) should be guided by the probable desirability of studying many sites, for reasons given in item 7 of section 2, General advice.

The primary consideration in choosing each geographic area is that it will yield legitimate scientific results. Without that, the efforts of the program are wasted. Included in the geographic choice are the type of aquatic habitat and the type of mine and waste discharge. The geographic variables which are most important for valid results will depend to some extent on the question that is being asked, but the following qualities are likely to be desirable at a site.

> The physical nature of the waterbody is such that it supports a balanced and full community of aquatic organisms.
>
> The physical nature of the waterbody is such that it can be sampled with precision and relative ease using standard techniques, particularly for benthic organisms.
>
> The mine discharge to be studied is relatively constant in nature so that it can be adequately sampled and reasonably characterized. At least it should not fluctuate wildly.
>
> There should be a good comparable control area ("upstream") for communitysampling. Without this the study loses almost all its power[9].
>
> The wastewater must not only cause a sublethal effect in the toxicity tests, but it must

---

[8] As is well known, the ordinary variables of surface water such as hardness, alkalinity, pH and temperature can have profound effects on toxicity of some substances, especially metals that might be present in mining wastes. There could be other modifying agents in the water of the study site, such as organic molecules that would bind and detoxify metals. There could be additional toxicants coming from upstream sources, be they known or unknown. Using upstream water includes all the above possibilities in the toxicity tests and makes them comparable to conditions downstream of the discharge, i.e. more relevant than would be the case with clean "laboratory" dilution water.

[9] As mentioned in section 2, there are statistical complications in an upstream/downstream study since theoretically, the communities are not independent. Nevertheless, it is usually better to have the control and experimental stations in the same waterbody because a control in an adjacent waterbody is likely to have greater differences in natural characteristics. As always, a statistician should be involved from the design stage.

also cause a downstream effect on the receiving community. Without both effects, a field validation will be inconclusive[10].

There should be a gradation of effects in the receiving water, from stronger effects in the plume to weaker effects with downstream dilution (tributaries), and finally (ideally) a disappearance of significant effect. That range would allow a strong comparison between degrees of effect in the community and the toxicity tests, at given concentrations. Sampling sites should be selected to span the useful range of concentrations. A gradient is important to strengthen the evidence that an effect of *pollution* is being measured, not some normal upstream/downstream phenomenon (see discussion under item 7 of General Advice, section 2).

The toxic components in the waste should be generally known, so that they can be adequately measured by chemical tests.

There should be a minimum of extraneous conflicting influences on the aquatic community, such as forestry-related impact, agricultural runoff, and upstream dams with unusual flow regimes.

Other significant geographic factors that are of secondary importance from a scientific point of view would include the following items.

No problems of accessibility because of conflicting property rights.

Reasonable access to the home laboratory. While shipping of samples and travel of personnel can be an economic millstone and must be considered, the scientific requirements are the first consideration in picking locations.

Phase 3. Survey existing records

Records at a particular site would include maps, history of the region (particularly industrial), and physical, chemical, and biological information on the waterbody.

It might be tedious to search for some of these, but collecting information anew would entail much more expenditure of effort and delay. For some information, such as seasonal variation in flow patterns, there is no choice except to obtain it from existing records.

---

[10] If no effect is documented, there is no measure of the toxic strength of the wastewater. A toxic effect might have been caused at a slightly higher concentration, or it might have required a much higher concentration. The attempted measurements would fail to provide the information desired, and effort and resources would have been wasted. The problem was seen in the site chosen for the pilot field study of AETE in 1995 -- no effect was demonstrated in the biological surveys. The AETE program to evaluate lethal tests in the laboratory had similar problems. Many of the effluents chosen for study did not cause an effect with one or more of the organisms, so no comparative quantitative information was obtained.

Phase 4. Initial screening and sampling program

The main variables in a desired validation study should be assembled into an rough plan, to assess the feasibility. Some information might be available from past records, but some might require exploratory sampling and documentation. Among the items which will require approximations are the qualities, magnitudes, and variations of wastewater toxicity, plume size and dynamics, and composition of the biological community. What range of concentrations will be needed in toxicity testing, to successfully obtain mild and strong effects? What kinds and numbers of invertebrates will be obtained in samples, and is the upstream community comparable to an unpolluted downstream community?

Does the rough plan make the validation program seem feasible? Are problems foreseen for some topics?

Phase 5. Preliminary modelling/design/analysis

This phase actually progresses in concert with phases 1 to 4. It is listed separately, but anyone setting up a scientific validation program would automatically set up an informal plan, then repeatedly rethink the design[11]. The details would probably change as more was learned about the site and industrial operations.

This phase should map out the general design of the program, methods of collection, analysis, and reporting.

Phase 6. Design of detailed validation program

The numerical values attached to items like those mentioned under phase 4, and especially their variation, should be used in designing the final program of sampling. At this stage there must be determination of the precision required in various parts of the program, and hence the numbers of replicates.

Several of the following items should be considered at this stage, and perhaps be included in the design of the investigative program, depending on the desired rigour.

(Item 6a) Scientific method

This is the time to make full use of the scientific method, although earlier planning (Phase 5) must also have paid attention to hypothesis-formulation and the information required.

Ask the question. What is the purpose of this work?

Formulate an hypothesis or hypotheses.

---

[11] It would be more important to separate this preliminary planning as a discrete phase, if one were designing a case of risk management analysis, or a watershed management program.

The hypotheses must be formally written down before the work is finally planned.
E.g. H1: The community of organisms in the receiving-water will be deleteriously affected at concentrations of mine-waste that cause sublethal effects in the toxicity tests.
E.g. H2: The community of organisms will not be affected at concentrations of mine-waste that do not cause sublethal effects in toxicity tests[12].

Design the simplest possible experiment that will answer the question. Consider the variation of components estimated in earlier stages as well as the usual variation in such work, and thus the magnitude and the power of the experiment needed.

Set up, ahead of time and formally, the criteria to judge the results, criteria by which it will be possible to accept or reject the hypotheses.
E.g. "The downstream aquatic community will be considered as affected in a deleterious manner if there is a statistical decrease at the 5% level of significance, compared to the control community, in [decide which] all of/any of/two of the Shannon diversity index, species richness, biotic index ..."
E.g. "In the sublethal toxicity test with *Ceriodaphnia*, a concentration of mine-waste will be considered toxic if it is equal to the IC25, or higher for [decide which] one of/both of the number of young produced, and/or the mortality among the first generation."
E.g. "In the sublethal test with Microtox (chronic) toxicity will be considered to exist if ..."

Formally set up, ahead of time, the statistical testing program that willbe used. This is an imperative for designing the whole program. Biologists have, in the past, been notorious for doing things in the very ineffective opposite way -- collecting a bunch of information, then casting about to see if there are any statistical tests which can be used to sort out the mess of data.

There could be secondary purposes for a validation study. As mentioned above, it might be desired to see whether chemical tests on the effluent would adequately predict the effects of a mine discharge. That would require setting up a formal program of physico-chemical testing as an addition to the program of toxicity testing and community assessment. In fact, that would seem a logical and useful adjunct study. To carry it out, the same phases should be considered, and separate hypotheses, criteria, and procedures should be set up. Costs would generally be only the add-ons to the primary biological program.

(Item 6b) Sensitivity analysis

This technique requires someone capable in modelling and familiar with sensitivity analysis. It is not commonly seen in water pollution studies, but could help avoid indeterminate results. It would most conveniently involve a model or models, preferably on computer, into which hypothetical results could be entered. One might enter, for example, various values for variance of the toxicity tests and see how each value affected statistical tests, the conclusions from the program, and indeed the ability to draw any conclusions.

---

[12] H1 and H2 are the reverse of each other, of course. It might not seem necessary to state them both or test them both, but it is necessary to have both answers for a complete study. It would not be very satisfying to find that a toxic concentration of effluent caused an effect in the community, but that a concentration which was non-toxic in the lab still caused community effects.

If it were shown that the overall results were particularly sensitive to any element of the program, say the sample-to-sample variation in collection of invertebrates, the program could be redesigned to increase the sampling effort, to change sampling apparatus, or to decrease the variance in some other way. Alternatively, the sensitivity analysis might indicate that some part of the program was over-designed compared the what was needed, so that some effort could be saved by re-design.

(Item 6c) Logistical support plan

This item would also be considered automatically in planning a program, but it deserves specific mention since it drives the economic feasibility of any plan. Who will do what, and when will it occur? What equipment is needed, when, and how will it get to the right places? Which laboratories will carry out tests, and on what schedules? How will the sampling personnel reach the locations, and the samples get to the laboratory? Costs can be estimated at this stage.

It is possible that completing the logistical plan will show that an adequate program will be more expensive than expected. Item d (below) should be considered and that item becomes much more important. The planners might examine their objectives and consider whether the essential questions could be answered by a smaller program and more focused objective. There could be managerial decisions to adopt existing background data for part of the program, or to reduce the scope of sampling. Such changes should be recycled through the preceding parts of Phase 6 to see the interaction with criteria for accepting the hypotheses, and the likelihood of reliable conclusions.

It might be that sufficient funding and resources could not be gathered to carry out an adequate and definitive program. In that case, rather than mount a reduced program that would probably be inconclusive and futile, *the program should be abandoned* (Sanders 1985).

(Item 6d) Cost-benefit analysis

This item is seldom carried out formally in water pollution work, but is always present as an unstated consideration which shapes programs and plans. For example there is always a decision on how much funding to provide for an investigative program. It would be desirable to formalize a cost-benefit analysis at this point, even with a simple analysis.

If the cost of the investigative program outweighed the environmental and economic benefits that might be achieved, obviously it might be a sensible decision to*abandon the program*. For example, the objective of the program being considered here is to validate sublethal toxicity tests. The implied reason for validation is that toxicity tests could be an additional technique for monitoring wastewater discharge, or perhaps a superior or cheaper technique, compared to chemical measurements or receiving-water studies.

A cost-benefit analysis would be best done by someone experienced in the technique, and with capability as an economist [13]. For example, if toxicity testing appeared to reduce monitoring costs, a qualified person could make comparisons with potential savings over the years for Canadian mines or categories of mines.

Phase 7. Carry out the program

Collect the samples, do the tests, and analyze the results, as planned. Are the hypotheses accepted or rejected?

The final stage is reporting the results. Sanders (1985) warns that this stage is sometimes difficult to complete because of cost over-runs during data-collection, because personnel lose interest upon completion of technical tasks, or go on to other projects, etc. Contract payments based on "milestone" achievements would seem a useful remedy.

4 Sources of information on planning field validation

Some of the superior studies reviewed in chapter 2 would be worth examining as potential patterns.

A chapter on design for field validation gives advice that is parallel to that given here, although more general (Sanders 1985). Many items from that chapter have been incorporated into this chapter. Three papers by Cairns and colleagues, given in the reference list, provide thoughtful orientation. Marcus and McDonald (1992) review eight studies of the U.S. EPA Program on Testing Toxicity of Complex Effluents, outlining proper methods of statistical comparison, and elucidating the differences between rigorous field validation and lower-level correlation studies. OECD (1992) presents good principles and recommendations derived from a workshop. Livingston and Meeter (1985) provide an academic focus on criteria for verification of laboratory and field results.

References

Berkman, H.E., C.F. Rabeni and T.P. Boyle 1986. Biomonitors of stream quality in agricultural areas: fish versus invertebrates. Environ. Management 10: 413-419.

---

[13] The economist should be environmentally sensitive. There still exist, people who would assign a value of zero to an ecosystem if it did not produce saleable goods.

Cairns, J. Jr. (ed.) 1982. Artificial substrates. Ann Arbor Science Pub. Inc., Ann Arbor, Michigan. 279 p.

Cairns, J. Jr. 1986. What is meant by validation of predictions based on laboratory toxicity tests? Hydrobiologia 137: 271-278.

Cairns, J. Jr. 1988a. What constitutes field validation of predictions based on laboratory evidence? P. 361-368 *in*: W.J. Adams, G.A. Chapman and W.G. Landis (eds). Aquatic Toxicology and Hazard Assessment: Tenth Volume. Amer. Soc. Testing and Materials, Philadelphia, Pa., ASTM STP 971.

Cairns, J. Jr., E.P. Smith and D. Orvos 1988. The problem of validating simulation of hazardous exposure in natural systems. Proc. 1988 Summer Computer Simulation Conf., p. 448-454. C.C. Barnett and W.M. Holmes (eds). Soc. Computer Simulation Internat., San Diego, Calif.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. Wiley-Interscience, Wiley & Sons, New York. 257 p.

Green, R.H. 1990. Power analysis and practical strategies for environmental monitoring. Environ. Res. 50: 195-205.

Hellawell, J.M. 1978. Biological surveillance of rivers. A biological monitoring handbook. Water Research Centre, Elder Way, Stevenage, Herts., SG1 1TH, England. 332 p.

Kaesler, R.L., J. Cairns Jr., and J.S. Crossman 1974. Redundancy in data from stream surveys. Water Res. 8: 637-642.

Karr, J.A. 1993. Defining and assessing ecological integrity: beyond water quality. Environ. Toxicol. Chem. 12: 1521-1531.

Livingston, R.J. and D.A. Meeter 1985. Correspondence of laboratory and field results: what are the criteria for verification? P. 76-88 *in*: Cairns, J. Jr. (ed.). Multispecies toxicity testing. Pergamon Press, New York, N.Y..

Marcus, M.D. and L.L. McDonald 1992. Evaluating the statistical bases for relating receiving water impacts to effluent and ambient toxicities. Environ. Toxicol. Chem. 11: 1389-1402.

NRA 1991. Proposals for statutory water quality objectives. Report of the National Rivers Authority. National Rivers Authority, Rivers House, Waterside Drive, Aztec West, Almondsbury, Bristol BS12 4UD, United Kingdom. NRA Water Quality Series No. 5, 99 p.

OECD (Organization for Economic Co-operation and Development) 1992. Report of the OECD workshop on the extrapolation of laboratory aquatic toxicity data to the real environment. OECD Environment Monographs No. 59. Paris, France. OCDE/GD(92)169.

Roch, M. and J.A. McCarter 1984b. Hepatic metallothionein production and resistance to heavy metals by rainbow trout *(Salmo gairdneri)* -- II. Held in a series of contaminated lakes. Comp. Biochem. Physiol. 77C: 77-82.

Sanders, W.M. III 1985. Field validation. P. 501-618 *in:* G.M. Rand and S.R. Petrocelli (eds). Fundamentals of aquatic toxicology. Methods and applications. Hemisphere Pub. Corp., Washington, D.C.

Sprague, J.B. 1986. Toxicity and tissue concentrations of lead, zinc, and cadmium for marine molluscs and crustaceans. Internat. Lead Zinc Research Organization Inc., Boca Raton, Florida. 215 p.

Washington, H.G. 1984. Diversity, biotic and similarity indices. A review with special relevance to aquatic ecosystems. Water Res. 18: 653-694.